

**ANA CATARINA JACINTA FERNANDES**

**THE ROLE OF miRNA-MEDIATED C/S-REGULATION IN  
BREAST CANCER SUSCEPTIBILITY**



**UNIVERSIDADE DO ALGARVE**

Departamento de Ciências Biomédicas e Medicina

2017



**ANA CATARINA JACINTA FERNANDES**

**THE ROLE OF miRNA-MEDIATED C/S-REGULATION IN  
BREAST CANCER SUSCEPTIBILITY**

**Master in Oncobiology – Molecular Mechanisms of Cancer**

**This work was done under the supervision of**

**Ana Teresa Maia, Ph.D**

**Joana Xavier, Ph.D**



**UNIVERSIDADE DO ALGARVE**

Departamento de Ciências Biomédicas e Medicina

2017



**THE ROLE OF miRNA-MEDIATED C/S-REGULATION IN  
BREAST CANCER SUSCEPTIBILITY**

**Declaração de autoria de trabalho**

Declaro ser a autora deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Ana Fernandes

(Ana Fernandes)

Copyright © 2017 Ana Catarina Jacinta Fernandes

A Universidade do Algarve reserva para si o direito, em conformidade com o disposto no Código do Direito de Autor e dos Direitos Conexos, de arquivar, reproduzir e publicar a obra, independentemente do meio utilizado, bem como de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição para fins meramente educacionais ou de investigação e não comerciais, conquanto seja dado o devido crédito ao autor e editor respetivos.

*If you are not willing to risk the unusual, you will have to  
settle for the ordinary.*

Jim Rohn





## ACKNOWLEDGMENTS

Em primeiro lugar, gostaria de agradecer à minha orientadora, Dra. Ana Teresa Maia, por me ter aceite no seu laboratório no *Centre for Biomedical Research* (CBMR), sediado na Universidade do Algarve, por me ter dado a oportunidade de acompanhar o seu grupo, por todo o conhecimento partilhado, e sobretudo por toda a compreensão, apoio e motivação prestadas. Obrigado por ter acreditado em mim.

Gostaria também de agradecer à minha co-orientadora, Dra. Joana Xavier, por todo o apoio, motivação e companheirismo. Não seria a mesma coisa sem alguém para nos “iluminar” a todos.

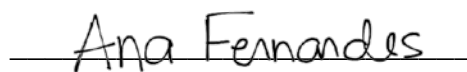
Obrigado ainda ao meu colega do lado, Ramiro Magno, por toda a ajuda e sabedoria. Espero que agora já saibas o meu nome completo, e e-mail de “cor e salteado”. Obrigado Filipa Esteves, por toda a paciência, disponibilidade e compreensão. Aprendi imenso com os dois.

À Mariana Jordão e Teresa Lourenço, gostaria de expressar a minha gratidão. Fomos sem dúvida feitas para nos “aturar” umas às outras.

Obrigada aos meus “amiguinhos” de licenciatura e mestrado, Bernardo Almeida, Ricardo Matos, Joel Lage e Pedro Charlito. Somos um grupo imbatível.

Às minhas amigas de infância, Joana Veiga, Eva Viegas, Filipa Ferreira e Tânia Rodrigues, agradeço-vos do fundo do coração. Obrigada por todo o apoio e incentivo durante estes últimos 13 anos. Será, sem dúvida, para a vida.

Finalmente, um agradecimento muito especial à minha mãe e irmão, Maria Zélia Guerreiro Jacinta e Ricardo Manuel Jacinta Fernandes, e a toda a minha família por todo o apoio incondicional. E claro, obrigada “Mika”, por estares sempre presente quando preciso de ti e por me fazeres rir todos os dias.



(Ana Fernandes)



## ABSTRACT

*Cis*-regulation of gene expression is believed to be central in breast cancer (BC) predisposition. Here we aimed to unravel the contribution of allele-specific miRNA regulation to BC risk.

We screened the effect of 223 published BC genome wide association studies (GWAS) -significant single nucleotide polymorphisms (SNPs) (and their 2668 unique proxies in high linkage disequilibrium) on differential miRNA-regulation. We filtered these SNPs based on location in miRNA genes and/or messenger RNA (mRNA) of protein-coding genes. Selected SNPs were then evaluated for putative differential miRNA-binding using TargetScan and miRanda, two distinct miRNA-target prediction algorithms, modified to analyse sequences carrying SNP alleles. Results were filtered for miRNAs with evidence of expression in breast tissue, and for genes displaying differential allelic expression (DAE), a hallmark of *cis*-regulation. To validate our findings, we prioritized the candidate SNPs for functional characterization, by combining TargetScan' and miRanda' predictions.

Interestingly, none of the SNPs mapped to miRNA genes, thus suggesting that miRNA biogenesis and target-binding alteration, via seed sequence modification, are mechanisms unlikely to be involved in BC risk. Of the SNPs located in mRNA sequences we found 93 out of 3891 that were predicted to alter the miRNA-mRNA binding in 27 BC-associated risk *loci*. From our predictions, we found rs4245739 in *MDM4* and rs11540855 in *ABHD8*, already functionally validated by others to cause allele-specific miRNA-binding.

We carried *in vitro* functional characterization of rs6884232 in *ATG10*, one of the best candidates identified by both TargetScan and miRanda algorithms. The predicted specific binding of hsa-miR-21-3p to the G allele of this SNP was evaluated using a dual-luciferase system, with constructs carrying either the A or the G allele, and in combination with miRNA mimics and inhibitors in a breast adenocarcinoma cell line. However, no allele-specific differences in luciferase activity were observed.

To our knowledge, this is the first study looking into the global role of miRNA regulation in BC risk, further improved by the integration of DAE data from normal breast samples.

#### Keywords

breast cancer • GWAS • *cis*-regulation • SNPs • miRNA

## RESUMO

A cis-regulação é um dos mecanismos pelo qual a expressão génica é maioritariamente controlada. Para além disso, postula-se que a cis-regulação tenha um papel fundamental para o risco de doenças complexas, onde se inclui o cancro da mama. Sendo o cancro da mama o tipo de cancro mais comum entre mulheres a nível mundial, o benefício em identificar marcadores de risco e de os usar na clínica, para a identificação precoce da população em risco, é indiscutível.

Nos últimos dez anos, estudos de associação genómica (do inglês *genome-wide association studies*, GWAS) têm vindo a identificar um largo número de variantes genéticas comuns que conferem baixos níveis de risco para cancro da mama. Estas variantes são polimorfismos de nucleótido único (SNPs) e estão maioritariamente localizadas em regiões não codificantes, o que sugere que estas poderão conferir risco através da regulação dos níveis de expressão génica. Dos estudos funcionais já efetuados para algumas destas variantes de risco, confirmou-se que estas são cis-reguladoras e que conferem risco para cancro da mama através da ligação diferencial de fatores de transcrição. No entanto, existem muitos outros mecanismos biológicos de cis-regulação para além da ligação de fatores de transcrição, tais como a regulação pós-transcricional por microARNs (miARNs) ou alteração do processamento do ARN. Particularmente, os miARNs são pequenas moléculas de ARN de cadeia simples, capazes de induzirem o silenciamento de genes alvo ao se ligarem por complementaridade, principalmente, à região 3' não traduzida (UTR) do ARN mensageiro. Os miARNs podem também levar ao silenciamento de genes ao ligar-se à sua sequência codificante (CDS) ou 5'UTR. Cis-regulação via miARNs pode ser afectada por SNPs localizados em genes codificantes para miARNs, afetando a sua biogénese ou alterando os seus genes alvo (através da alteração da sequência de ligação dos miARNs), ou podem estar localizadas nos genes alvo, alterando a estabilidade de ligação dos miARNs.

A presente dissertação de mestrado teve como objetivo desvendar a contribuição de variantes genéticas cis-reguladoras que afetam a regulação de miARNs para o risco de cancro da mama.

Como tal, investigou-se o efeito de 223 SNPs, identificados por GWAS para risco para cancro da mama, na regulação diferencial por miARNs. Da mesma forma, também se avaliou o efeito de SNPs que se encontravam em elevado desequilíbrio de ligação com estes, ou seja em estrita ligação genética. Primeiro, filtrou-se estes SNPs pela sua localização em genes de miARNs e/ou 5'UTR, CDS e 3'UTRs de genes codificantes para proteínas. Posteriormente, avaliou-se o potencial dos SNPs selecionados para alterarem a ligação de miARNs. Uma vez que não se encontravam disponíveis ferramentas que o efetuassem, procedeu-se à modificação de dois algoritmos de previsão de ligação de miARNs já existentes, TargetScan e miRanda, para que estes permitissem a análise dos diferentes alelos de SNPs. Estes algoritmos foram selecionados não só pela sua metodologia, como também pela sua capacidade de serem aplicados em R, um ambiente de programação de livre-acesso para computação estatística e gráfica. Deste modo, foi possível não só modificação dos algoritmos para analisarem sequências contendo alelos de SNPs, como também foi possível a sistematização do processo.

De seguida, efetuou-se um filtro para miARNs expressos em tecido mamário, por estes possuírem expressão específica de tecidos. Para além disso, filtraram-se os resultados por genes que possuísssem expressão alélica diferencial (*DAE*), uma característica da *cis*-regulação, em tecido mamário normal. De modo a validar as nossas descobertas, procedeu-se à priorização dos SNPs candidatos para validação funcional. Para isso, combinaram-se as previsões efetuadas tanto pelo *TargetScan* como pelo *miRanda*, de modo a aumentar a probabilidade de selecionar um SNP que tivesse um efeito real.

Curiosamente, nenhum dos *SNPs* associados com risco para cancro da mama se encontrava em genes de miARNs. Isto sugere que tanto a alteração da biogénese de miARNs, como a alteração dos genes alvo por modificação da sequência de ligação dos miARNs, são mecanismos improváveis de estarem a contribuir para o risco para cancro da mama.

Dos SNPs localizados em genes codificantes para proteínas, encontraram-se 93 (dos 3891 SNPs iniciais) previstos de estarem a afetar a ligação de miARNs em 27 *loci* associados com risco para cancro da mama. Isto sugere que cerca de um quarto dos *loci* já associados com risco para cancro da

mama podem ser explicados pela regulação diferencial por miARNs. Destes SNPs, dois já se encontram validados funcionalmente noutros estudos, como estando a causar a ligação específica de miARNs para diferentes alelos: rs4245739 no gene *MDM4* e rs11540855 no gene *ABHD8*. Isto valida a nossa análise e sugere que outras das previsões efetuadas também poderão ser funcionais.

Finalmente, a priorização dos SNPs de risco, associados com cancro da mama, efetuada através da combinação das previsões obtidas tanto pelo TargetScan, como pelo miRanda, resultou na identificação de seis candidatos com maior probabilidade de estarem a afetar a ligação de miARNs: rs1573 (localizado no gene *ASB13*), rs2385088 (localizado no gene *ISYNA1*), rs1019806 e rs6884232 (localizados no gene *ATG10*), rs4808616 (localizado no gene *ABHD8*), e ainda rs3734805 (localizado no gene *CCDC170*).

De seguida, procedeu-se à caracterização funcional *in vitro* do rs6884232 localizado no gene *ATG10*. Para este SNP, previu-se a ligação específica do hsa-miR-21-3p ao alelo G (*context++ score* = -0.169), o que resultaria na diminuição da expressão deste gene. Primeiro, efetuou-se um ensaio de luciferase em células de adenocarcinoma mamário (*MCF-7*) usando plasmídeos com genes repórter de luciferase contendo, ou o alelo A, ou o alelo G do SNP. No entanto, não foram observadas diferenças na atividade da luciferase entre ambos os alelos. De seguida repetiu-se o ensaio, desta vez em combinação com cósmicos e inibidores do hsa-miR-21-3p e ainda com os seus respetivos controlos negativos. Mais uma vez, não se obtiveram diferenças na atividade da luciferase entre ambos os alelos, sugerindo que este SNP não causa a ligação diferencial do hsa-miR-21-3p aos seus alelos. Porém, a elevada variabilidade obtida entre replicados biológicos, assim como efeitos não esperados em condições controlo, não nos permite ainda retirar conclusões definitivas, sendo necessário repetir o ensaio.

Tanto quanto se sabe, o presente trabalho é o primeiro estudo a avaliar o papel global da regulação por miARNs no risco para cancro da mama e que engloba dados de *DAE* em tecido mamário normal. No futuro, esperamos complementar esta abordagem ao determinar e caracterizar a importância clínica do efeito de variantes genéticas cis-reguladoras mediadas por miARNs

em cancro da mama. Isto permitirá melhorar a caracterização dos *loci* já associados com risco para cancro da mama e ainda melhorar o conhecimento da etiologia do cancro da mama.

#### Palavras-Chave

cancro da mama • GWAS • *cis*-regulação • SNPs • miRNA



# CONTENT

<b>Acknowledgments</b> .....	vii
<b>Abstract</b> .....	ix
<b>Resumo</b> .....	xi
<b>Content</b> .....	xv
<b>Index of Illustrations</b> .....	xix
<b>Index of Tables</b> .....	xxi
<b>Index of Annexes</b> .....	xxii
<b>List of Abbreviations</b> .....	xxiii
<b>1. Introduction</b> .....	3
1.1 Genetic variation in gene expression levels .....	4
1.1.1 Expression quantitative trait loci (eQTL) mapping .....	5
1.1.2 Differential allelic expression (DAE) .....	7
1.2 Molecular mechanisms of cis-acting genetic variation .....	9
1.3 Genetic variation in common complex traits and diseases .....	11
1.4 microRNA .....	12
1.4.1 miRNA biogenesis .....	12
1.4.2 miRNA targeting .....	15
1.4.3 miRNAs in common complex diseases .....	17
1.5 Breast Cancer .....	18
1.5.1 Epidemiology .....	18
1.5.2 Aetiology .....	19
1.5.3 Breast Cancer Familial Risk .....	20
1.5.3.1 High-risk mutations .....	21
1.5.3.2 Moderate-risk mutations .....	22
1.5.3.3 Common modest-risk alleles .....	22
1.5.3.3.1 GWAS .....	23

1.6	Prioritizing cis-regulatory candidates.....	25
1.7	Preliminary data.....	26
<b>2.</b>	<b>Aim .....</b>	<b>27</b>
<b>3.</b>	<b>Materials and Methods .....</b>	<b>31</b>
3.1	Dataset .....	31
3.2	miRNA-target prediction.....	31
3.2.1	R programming language and RStudio® IDE .....	31
3.2.2	SNP query with SNAP (Broad Institute) .....	32
3.2.3	miRNA-target binding prediction algorithms.....	33
3.2.3.1	TargetScan .....	33
3.2.3.1.1	Tool Description .....	33
3.2.3.1.2	miRNA::mRNA interaction predictions.....	33
3.2.3.2	miRanda.....	35
3.2.3.2.1	Tool Description .....	35
3.2.3.2.2	miRNA::mRNA interaction predictions.....	35
3.2.4	miRNA expression in breast .....	37
3.2.5	Plot generation .....	37
3.2.6	DAE SNP ranking.....	38
3.3	Functional validation .....	38
3.3.1	Luciferase Reporter Gene Assays .....	38
3.3.1.1	Oligonucleotide design and plasmid selection .....	39
3.3.1.2	Annealing, digestion and ligation.....	40
3.3.1.3	Bacterial transformation .....	42
3.3.1.4	Plasmid amplification, extraction and purification .....	43
3.3.1.5	Cell line .....	43
3.3.1.6	Cell culture .....	44
3.3.1.7	Transfection and luciferase assay .....	44

3.3.1.8	Statistical analysis .....	46
<b>4.</b>	<b>Results .....</b>	<b>49</b>
4.1	Dataset of breast cancer risk-associated SNPs.....	49
4.2	miRNA biogenesis and seed region alteration are mechanisms unlikely involved in breast cancer risk.....	50
4.3	Most SNPs associated with breast cancer risk are located to the non- coding regions of protein coding genes.....	51
4.4	Allele-specific miRNA target-prediction analysis .....	51
4.4.1	TargetScan analysis .....	53
4.4.2	miRanda analysis .....	54
4.4.3	Candidate prioritization .....	55
4.5	Functional Validation of rs6884232 differential allelic binding to hsa- mir-21-3p .....	60
<b>5.</b>	<b>Discussion .....</b>	<b>67</b>
5.1	Localization of BC risk SNPs .....	67
5.2	Comparison of miRNA-target prediction algorithms.....	68
5.3	Expression and cis-regulation in breast.....	69
5.4	SNP candidate prioritization for functional validation.....	71
5.5	rs6884232: a cis-regulatory candidate .....	73
5.5.1	Functional validation of rs6884232 .....	74
<b>6.</b>	<b>Conclusions and Future Directions .....</b>	<b>79</b>
<b>7.</b>	<b>References .....</b>	<b>80</b>
	<b>Annexes .....</b>	<b>98</b>



## INDEX OF ILLUSTRATIONS

Figure 1.1 Example of a cis- and/or trans-regulatory effect by a non-coding genetic variant.....	5
Figure 1.2 Representation of a typical eQTL. ....	6
Figure 1.3 The advantage of DAE analysis compared to eQTL mapping in the identification of cis-regulation. ....	8
Figure 1.4 Overview of cis-regulatory mechanisms. ....	10
Figure 1.5 miRNA gene types and canonical biogenesis.....	15
Figure 1.6 Canonical and non-canonical miRNA binding.....	16
Figure 1.7 Globocan 2012 illustrating the world's breast cancer burden.....	18
Figure 1.8 Breast cancer susceptibility loci by risk allele frequency and conferred relative risk.....	20
Figure 1.9 Schematic diagram of the haplotype blocks in two populations and relation to GWAS. ....	24
Figure 3.1 Flowchart of differential miRNA-target prediction analysis.....	36
Figure 3.2 Mechanism of action of a luciferase reporter gene assay developed to test miRNA-mediated post-transcriptional regulation of target genes. ....	39
Figure 3.3 pmiRGLO vector and oligonucleotide inserts.....	41
Figure 3.4 E. coli transformation protocol. ....	42
Figure 3.5 MCF-7 cell line. (Adapted from ATCC) .....	44
Figure 3.6 Luciferase assay transfection protocol. ....	46
Figure 4.1 Chromosomal localizations and transcript consequences of BC GWAS-significant SNPs. ....	49
Figure 4.2 Venn diagram containing the genomic localizations of the 3709 queried SNPs.....	50
Figure 4.3 Venn diagram regarding the location of the candidate cis-regulatory SNPs within protein-coding genes.....	51
Figure 4.4 Venn diagram showing the relation between the total number of predictions obtained from the miRanda and TargetScan algorithms for an initial universe of 45 SNPs located in 3'UTRs.....	55
Figure 4.5 Top six predictions. ....	59
Figure 4.6 DNA gel electrophoresis for pmirGLO linearization and ligation assessment.....	61

Figure 4.7 Transformation and insert integration confirmation.....	61
Figure 4.8 Dual-Glo Luciferase Assay for rs6884232. ....	62
Figure 4.9 Dual-Glo Luciferase Assay for rs6884232 and hsa-miR-21-3p. ....	63

**INDEX OF TABLES**

Table 3.1 Features included in the context++ score calculated by TargetScan.  
..... 34

Table 3.2 rs6884232 oligonucleotide pairs. .... 40

Table 4.1 Comparison between the freely-available web-based miRNA-target  
prediction algorithms. .... 52

Table 4.2 Common TargetScan and miRanda predictions for miRNAs  
expressed in normal breast tissue. .... 58

**INDEX OF ANNEXES**

Annex A. List of GWAS-significant SNPs. .... 98

Annex B. 3'UTR-located SNPs..... 99

Annex C. 54 unique SNPs located in the 5'UTR and/or CDS are listed.....102

Annex D. TargtetScan analysis pipeline. Online version only. ....104

Annex E. eQTL association of rs6884232 with ATG10 expression levels in breast tissue.....104



# LIST OF ABBREVIATIONS

## A

**ABHD8** – abhydrolase domain containing 8

**Ago** – Argonaut

**AI** – allelic imbalance

**Amp<sub>r</sub>** –  $\beta$ -lactamase gene

**APOC3** – apolipoprotein C3

**ASB13** – ankyrin repeat and SOCS box containing 13

**ATG10** – Autophagy related protein 10

**ATM** – ataxia-telangiectasia mutated

**ATP6AP1L** – ATPase, H<sup>+</sup> transporting, lysosomal accessory protein 1-like

**ATXN7** – ataxin 7

## B

**BC** – breast cancer

**bp** – base pair

**BRCA1/2** – breast cancer 1/2

**BRIP1** – BRCA1-interacting protein 1

## C

**CCDC170** – Coiled-Coil Domain Containing 170

**CD160** – natural killer cell receptor BY55

**CDS** – coding sequence

**CHECK2** – checkpoint kinase 2

**CREs** – cis-regulatory elements

**CYP51A1** – cytochrome P450 family 51 subfamily A member 1

## D

**DAE** – differential allelic expression

**DGCR8** – Di George critical region 8

**DLX2** – distal-less homeobox 2

**DMEM** – Dulbecco's modified eagle medium

**ds** – double-stranded

## E

**eQTL** – expression quantitative trait loci

**EXP5** – Exportin 5

## L

**luc2** – firefly luciferase reporter

## F

**FBS** – foetal bovine serum

**FGFR2** – fibroblast growth factor receptor 2

## G

**GWAS** – genome-wide association studies

## H

**HGK** – Human phosphoglycerate kinase

**hRluc-neo** – humanized Renilla luciferase-neomycin resistance cassette

**HSC70** – heat shock cognate 71 kDa

**HSP90** – heat shock protein 90

## I

**IDE** – integrated development environment

**ISYNA1** – inositol-3-phosphate synthase 1

## K

**kb** – kilobase

## L

**L1CAM** – L1 cell adhesion molecule

**LB** – Luria-Bertani

**LD** – linkage disequilibrium

## M

**MAF** – minor allele frequency

**MCF-7** – Michigan Cancer Foundation-7

**MCS** – multiple cloning site

**MDM4** – Human homolog of double minute 4, P53-Binding Protein

**METABRIC** – Molecular Taxonomy of Breast Cancer International Consortium

**MIER3** – mesoderm induction early response 1 family member 3

**MiRNA** – microRNA

**mRNA** – messenger RNA

## N

**ncRNA** – non-coding RNA

## P

**PALB2** – partner and localizer of BRCA2

**PAZ** – Piwi/Argonaute/Zwille

**PCGs** – protein-coding genes

**PEI** – Polyethylenimine

**pluc232A** – pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232

**pluc232G** – pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232

**pol** – polymerase

**pre-miRNA** – precursor miRNA

**pri-miRNA** – primary miRNA

**PSMD6** – proteasome 26S subunit,  
non-ATPase 6

## R

**RBD** – RNA-binding domain

**RIIID** – RNase III domains

**RISC** – RNA-induced silencing  
complex

**RNAi** – RNA interference

**RNase** – RNA ribonuclease

**RPM** – reads per mapped million

**RNF146** – ring finger protein 146

**RPS23** – ribosomal protein 23

## S

**s.e.m.** – standard error of mean

**SLC4A7** – solute carrier family 4  
member 7

**SNAP** – SNP Annotation and Proxy  
Search

**SNPs** – single nucleotide  
polymorphisms

**ss** – single-stranded

**SOC** – super optimal broth with  
catabolite repression

**STK11** – serine/threonine kinase 11

**stRNA** – small temporal RNAs

**SV40** – Simian Virus 40

## T

**TBE** – Tris-Borate-EDTA

**TERT** – telomerase reverse  
transcriptase

**TFs** – transcription factors

**TGFB1** – transforming growth factor  
beta 1

**TNIP3** – TNFAIP3 Interacting  
Protein 3

**TOMM20** – translocase of outer  
mitochondrial membrane 20

**TP53** – tumour protein 53

**TRBP** – Tar RNA binding protein

## U

**UTR** – untranslated regions

## W

**WT** – wild-type

## X

**XRCC1** – X-ray repair cross  
complementing 1



# **CHAPTER I**

## **Introduction**



## 1. INTRODUCTION

Phenotypic diversity is a result of environmental factors, epigenetic modifications and genetic variation modulating the expression of genes as well as changing the function of proteins. In the last two decades, several genetics and genomics studies have tried to unveil the proportion of total phenotypic variation that is due to genetics – heritability – and more importantly, to dissect the molecular mechanisms that cause it. Interestingly, measurements of steady-state messenger RNA (mRNA) levels have led to the observation that gene expression variation is quite extensive between unrelated individuals but similar within families (Cheung et al., 2003; Morley et al., 2004), indicative of a genetic contribution to the baseline expression of many genes.

Supported by microarray technology, expression profiling and genome-wide mapping studies provided formal evidence that strong heritable factors dictate differences in gene expression levels (Monks et al., 2004; Morley et al., 2004); and several genetic variants have been associated with common traits and diseases (MacArthur et al., 2017). However, the mechanisms by which they affect the phenotype remains poorly understood and its better characterisation is needed to improve targeted screening and personalised risk estimates that aim to preclude the occurrence of common complex diseases.

In the present chapter, I focus in normal gene expression variation as a major result of *cis*-acting genetic variation. Moreover, I discuss how several studies support the importance of *cis*-regulatory genetic variation in the human genome, whilst not providing information about the underlying regulatory mechanism. Additionally, from the possible *cis*-regulatory molecular mechanisms that cause allelic expression differences, I particularise the case of single nucleotide polymorphisms (SNPs) altering microRNA (miRNA) regulation. I also discuss the role of the latter in disease susceptibility, particularly focusing in breast cancer (BC) risk. Finally, I discuss the need for systematic studies focusing on the global role of specific regulatory mechanisms in BC risk assessment.

## 1.1 GENETIC VARIATION IN GENE EXPRESSION LEVELS

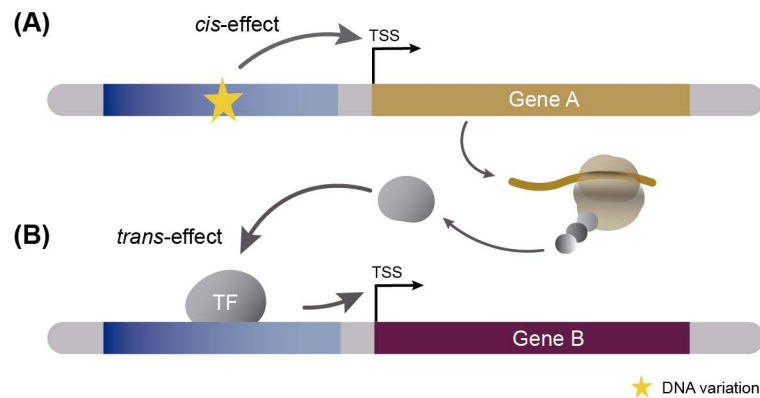
mRNA levels are controlled both by *cis*-acting elements and *trans*-acting factors (Pastinen and Hudson, 2004). *Cis*-acting elements or *cis*-regulatory elements (CREs) are DNA sequences that regulate gene expression (such as promoters and enhancers), as they often contain binding sites for other molecules (such as transcription factors, TFs). CREs are normally located near the gene they regulate and affect gene expression in an allele-specific manner. Conversely, *trans*-acting factors are molecules that regulate gene expression levels by binding to CREs. *Trans*-regulatory factors are normally encoded by genes located farther from the one(s) they regulate and can alter simultaneously both alleles of the target gene(s).

Accordingly, genetic variation, either in the form of mutations or polymorphisms, can lead to inter-individual variation in gene expression levels by acting in *cis*, if located in CREs, or in *trans*, if located in sequences that affect the activity or expression of *trans*-regulatory factors (Stranger et al., 2007a; Williams et al., 2007) – **Figure 1.1**.

*Cis*-acting genetic variants are therefore located in non-coding elements such as promoters, enhancers and introns of protein-coding genes (PCGs), possibly altering transcription, splicing, translation and stability, both at the mRNA and protein levels, in an allele-specific manner (Pastinen and Hudson, 2004; Williams et al., 2007). This gives rise to unequal allelic expression, a common feature in the human genome (Lo et al., 2003; Morley et al., 2004; Yan et al., 2002a).

*Trans*-acting genetic variants can either be located in the coding sequence (CDS) of PCGs, altering it, or be located in non-coding regions.





**Figure 1.1 Example of a *cis*- and/or *trans*-regulatory effect by a non-coding genetic variant.** (A) *Cis*-regulatory genetic variants are mainly, but not always (not shown), located near the gene they regulate (in this case, near gene A) and affect gene expression in an allele-specific manner. (B) If for instance, that gene codes for a transcription factor (TF), by affecting its expression, the genetic variant can also act in *trans* by affecting the expression of another, distally located, gene (gene B). This example also elucidates the fact that *cis*-acting genetic variations can also represent the first step in most *trans*-acting regulation. Transcription start site (TSS)

Determining the relative contribution of both these types of regulatory genetic variation to human gene expression, has been the aim of several studies in the last fifteen years and has largely focused on expression quantitative trait *loci* (eQTL) mapping (Cheung et al., 2003; Morley et al., 2004; Schadt et al., 2003; Stranger et al., 2007a).

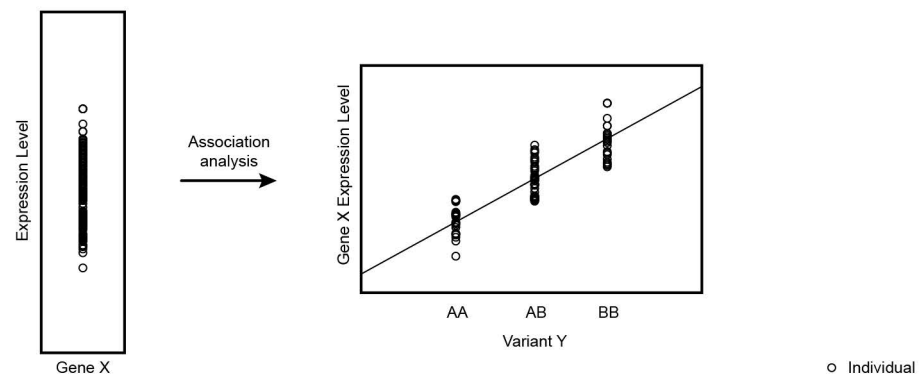
### 1.1.1 Expression quantitative trait *loci* (eQTL) mapping

In eQTL studies, total mRNA levels for a given gene are quantified and tested for association with specific genetic variation markers – **Figure 1.2**. These genetic variation markers are biallelic variants, for which three possible genotypes exist in the population. If different genotypes correspond to significantly different levels of gene expression, then the genetic variant is associated with the expression of a particular gene (Hulse and Cai, 2013) – eQTL.

eQTLs are often classified according to the relative locations (either the length of the physical or genetic distance) of the associated genetic variant and the regulated gene(s), and according to the type of mechanism through which they are thought to act. If the associated variant is located near the gene it

regulates, then it is termed as a “local” eQTL. Local eQTLs can act in *cis*, if they affect gene expression in an allele-specific manner – *cis*-eQTL – or, less commonly, act in *trans*, if they alter a trans-regulatory factor – *trans*-eQTL (Albert and Kruglyak, 2015; Cookson et al., 2009). On the other hand, if the associated genetic variant is distally located from the gene(s) it regulates, then it is referred to as a “distant” eQTL. Distant eQTLs normally act in *trans*.

Through the mapping of interindividual variation in gene expression levels, eQTL studies provided insightful information about the spatial distribution of regulatory variants in the genome (Veyrieras et al., 2008); and the level of steady-state expression variation that is associated with genetic variation in either proximal or distal regulatory elements (Cheung et al., 2003). This field is currently fostered by progressively advancing technologies and platforms that enable the systematic genotyping and gene expression measurements of increasing sample sizes, crucial to the reliable identification of variants with smaller impacts in gene expression levels (Cookson et al., 2009; Pai et al., 2015).



**Figure 1.2 Representation of a typical eQTL.** For a particular gene X with extensive gene expression variation among individuals, expression quantitative trait *loci* (eQTL) studies try to decompose that variation by associating it to the two alleles of a genetic variant. Thus, for a given biallelic genetic variant (variant Y), if to the three possible genotypes observed in the population (AA, AB, BB) correspond significantly different levels of gene X expression, then the genetic variant Y is associated with the levels of expression of gene X (Adapted from Cheung, 2016).

Heritability studies suggest that *trans*-acting genetic variation quantitatively accounts for most of the heritable gene expression differences (Price et al., 2011). However, the report of distant eQTL in humans remains challenging, mainly due to its small effect sizes (as compared to local eQTL;

Morley et al., 2004; Schadt et al., 2003) but also to higher statistical penalty required in multiple testing.

On the other hand, *cis*-acting genetic variation has been readily assessed by eQTL mapping and was suggested to account for 25-35% of interindividual gene expression differences (Morley et al., 2004; Schadt et al., 2003). Such characterisation has generated extensive controversy and likely represents an underestimate. In the presence of feedback mechanisms, subtle *cis*-regulatory variants can be overlooked by total mRNA measurements (performed in eQTL analysis); and several detected distant eQTLs are likely a consequence of *cis*-acting genetic variation (Pierce et al., 2014; Yang et al., 2016).

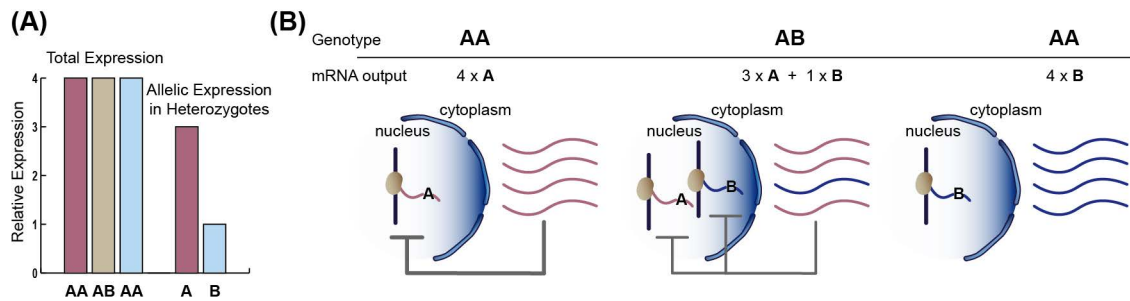
Importantly, eQTL studies strongly support the role of *cis*-regulatory genetic variation in the human genome (Aguet et al., 2016; Bryois et al., 2014; Göring et al., 2007; Stranger et al., 2007b). Also, they do not imply that *trans*-regulation of gene expression by genetic variants is uncommon or irrelevant, but that genetic variation with stronger or more detectable effects tend to be located/act in *cis*.

Nevertheless, gene expression is also affected by *trans*-regulatory factors and measurements of total mRNA levels in eQTL studies cannot discriminate the effect between *trans*-regulatory factors and CREs. Direct assessment of *cis*-regulation requires allele-specific approaches because the effect of *trans*-regulation is eliminated when comparing allelic expression in the same cell context (Pastinen and Hudson, 2004).

### 1.1.2 Differential allelic expression (DAE)

Allele-specific mRNA quantification provides the advantage of intrinsically controlling for both environmental and *trans*-regulatory factors affecting gene expression levels. These factors normally preclude the identification of subtle *cis*-acting genetic variations in other methodologies – **Figure 1.3**. By labelling the different alleles of transcribed genetic markers from a heterozygous

individual and subsequently quantifying the mRNA by-products, *cis*-regulation can be inferred when the allelic ratio is different from 1. Differential allelic expression (DAE; also known as allelic imbalance, AI) not only identifies *cis*-regulation, but also certain epigenetic effects, such as imprinting and random mono-allelic expression.



**Figure 1.3 The advantage of DAE analysis compared to eQTL mapping in the identification of *cis*-regulation.** In the presence of *trans*-acting feedback (in this case, negative feedback), measurements of total expression levels (performed in expression quantitative trait *loci*, eQTL) remain unchanged across both homozygotes and heterozygotes for a particular transcribed genetic variant. Although the A allele for a certain gene possesses higher expression activity due to an unidentified *cis*-acting genetic variant, AA homozygotes are strongly repressed by negative feedback mechanisms. As for the BB homozygotes, because the transcript containing the B allele has intrinsic lower expression levels, negative feedback is not active. AB heterozygotes have intermediate levels of shared negative feedback but its mRNA output still reflects their intrinsic expression levels derivative of differential *cis*-regulation. As a result, allele-specific approaches can be used to uncover *cis*-variation (Adapted from Pastinen, 2010).

DAE is a common feature in the human genome (Lo et al., 2003; Morley et al., 2004; Yan et al., 2002a) and both DAE and eQTL mapping studies strongly suggest a central role for *cis*-acting genetic variation in human gene expression.

However, despite the many benefits of eQTL mapping, which helped to systematically annotate several putative *cis*-regulatory *loci*, the mere mapping of a *locus* does not provide information about the affected biological mechanism, even when considering that the true regulatory causal variant was identified. In addition, even though DAE analysis provides a more direct assessment of *cis*-regulation, the causal variant remains unidentified, as well as the molecular mechanism that creates the allelic imbalance. Understanding this mechanism is crucial when dealing with common complex diseases, like BC

and type 2 diabetes, for which feasible and efficient targeted screening and prevention are decisive to reduce their overwhelming incidence.

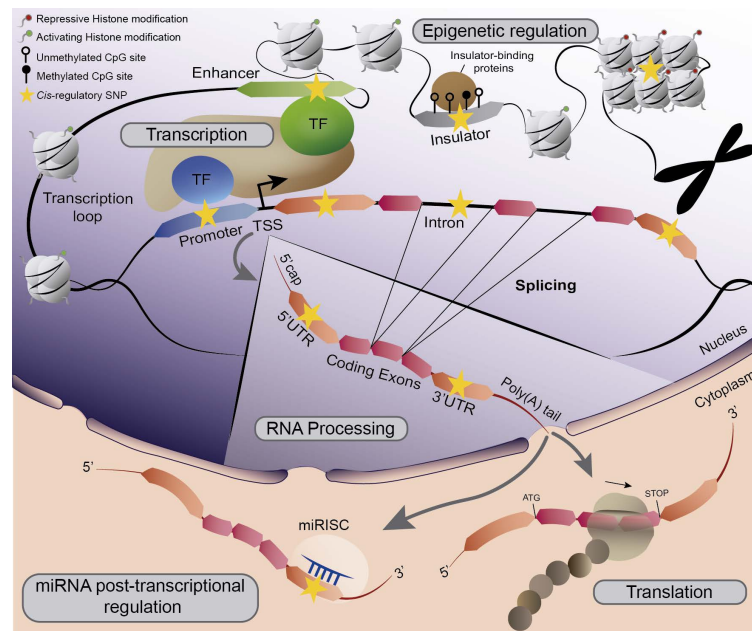
## 1.2 MOLECULAR MECHANISMS OF *CIS*-ACTING GENETIC VARIATION

CREs are non-coding DNA sequences that regulate gene expression. These sequences contain binding sites for other molecules, such as TFs. As a result, promoters and enhancers are the most well studied types of CREs (Huang and Ovcharenko, 2015; Sinnett et al., 2006; Wittkopp and Kalay, 2011). Nevertheless, several other types exist and albeit they are less understood, they may have considerable impact on gene expression. Ultimately, CREs are known to have a wide range of genomic locations, from distant enhancers/silencers and alternative promoters to downstream regulatory elements in untranslated regions (UTR) and introns. These elements act in a combinatorial manner to form regulatory programs that comprise a rather robust set of instructions capable of withstanding natural variations and environmental conditions.

If genetic diseases are the extreme manifestation of genetic variation, then normal variation in gene expression levels represents an intermediary state between genetic differences and complex human phenotypes, conferring different levels of risk. Deviations in gene expression levels, even small ones, can be important if the associated gene plays an important role in cellular pathways (Yan et al., 2002b).

In humans, interindividual genetic differences are mainly due to common variations in their genomic sequence, i.e., polymorphisms (Frazer et al., 2007; Sudmant et al., 2015). A polymorphism is defined as the existence of two or more genotypes in a population, given that the least frequent variant has an incidence equal or greater than 1%. By occurring nearly once in every 300 base pair (bp), SNPs are the most common type of genetic variation. This single bp substitution of genomic DNA is able to create, destroy or modify the binding sites found at CREs (Nicoloso et al., 2010; Philips et al., 2012; Sinnett et al.,

2006). **Figure 1.4** illustrates the overall landscape underlying the *cis*-effect of non-coding SNPs in the regulation of PCGs.



**Figure 1.4 Overview of *cis*-regulatory mechanisms.** Gene expression can be affected by *cis*-regulatory single nucleotide polymorphisms (SNPs), here represented by yellow stars, through several mechanisms. They can mediate their effect by altering transcription factor (TF) binding sites in proximal or distal relation of the transcription start site (TSS), such as those located in enhancers, silencers (not shown), insulators and promoter sequences. Also they can impair epigenetic regulation through the disruption of CpG sites. RNA processing mechanisms such as splicing and polyadenylation can be further altered by SNPs alleles disturbing conserved sequences in the 5' untranslated region (UTR), introns and 3'UTR of protein-coding genes. Additionally, *cis*-regulatory SNPs could also modify RNA post-transcriptional regulation by microRNA (miRNA), through the creation, modification or elimination of miRNA binding sites. Taken all together, these regulatory variants can modulate protein translation.

### 1.3 GENETIC VARIATION IN COMMON COMPLEX TRAITS AND DISEASES

Based on the assumption that common complex diseases are often caused by common variants (i.e., polymorphisms), called the “Common Disease, Common Variant” hypothesis, genome-wide association studies (GWAS) were broadly applied in the attempt to identify several risk *loci* responsible not only for complex diseases but complex traits, such as height or blood pressure (revisited in subchapter 1.5.3.3.1). Interestingly, the majority of these identified risk *loci* lie in non-coding regions (MacArthur et al., 2017), suggesting that they are regulating gene expression levels. In fact, for a few *loci*, a *cis*-regulatory role has been validated through functional studies.

Since there is a prevalent notion that transcriptional mechanisms might have a bigger toil on phenotypic expression and variation, most of the latter studies focus on the way SNPs can affect TF binding sites, overlooking other possible regulatory mechanisms, namely miRNA post-transcriptional regulation. By this mechanism, SNP alleles can either create, destroy or modify miRNA binding sites if located in the miRNA “seed” region or in the target-mRNA seed binding region. Also, they can regulate the expression levels of the miRNAs themselves or even their correct processing.

## 1.4 microRNA

miRNAs are endogenous single-stranded (ss) RNA molecules, about 22 nucleotides long, which are able to bind mRNA complementary sequences, either partial or fully, mainly in their 3'UTR but also in their CDS and 5'UTR.

They were initially discovered in the 90s as developmental timing regulators in the worm *Caenorhabditis elegans* (Lee et al., 1993; Ruvkun et al., 2000) and thus termed, at the time, small temporal RNAs (stRNA; Pasquinelli et al., 2000). The discovery of several new “tiny” non-coding RNA (ncRNA) in worm, fly and human, where most of which were not developmental stage related, shifted the nomenclature to miRNA in order to represent the previously described stRNA and other ncRNAs with similar features (Lagos-Quintana, 2001; Lau, 2001; Lee and Ambros, 2001).

Importantly, miRNA are thought to confer robustness to biological mechanisms by fine-tuning transcriptional programs (Ebert and Sharp, 2012). Particularly in humans, it is estimated that about 30% of PCGs are regulated by miRNAs, where each miRNA can target about 200 transcripts and each mRNA can be targeted by more than one miRNA simultaneously (Lewis et al., 2005).

### 1.4.1 miRNA biogenesis

Most miRNA genes are located in intergenic regions or introns of PCGs, predominantly in sense but also in antisense orientation (Kozomara and Griffiths-Jones, 2014). Currently, the database of miRNA sequences and annotation – miRBase (build 21) – contains 1881 precursors and 2588 mature human miRNA annotations (Kozomara and Griffiths-Jones, 2014), a number which by far surpasses the initial estimates of human miRNA genes (Lim, 2003). **Figure 1.5** illustrates several types of miRNA genes and the canonical miRNA biogenesis.

miRNAs are largely transcribed by RNA polymerase (pol) II into primary miRNA (pri-miRNA) transcripts. These transcripts are typically over 1 kilobase (kb) long and are composed of a 33-35 bp hairpin stem, a terminal loop and



ssRNA segments at the 5' and 3' ends, which are capped and polyadenylated (Cai et al., 2004; Lee et al., 2002, 2004) – shown in **Figure 1.5**.

pri-miRNAs are processed by the nuclear RNA ribonuclease (RNase) III Drosha and its partner Di George critical region 8 (DGCR8; Han et al., 2004; Lee et al., 2003). Together, Drosha and DGCR8 form a microprocessor which cleaves the pri-miRNA into a ~60-70 nucleotide precursor miRNA (pre-miRNA).

Drosha has a double-stranded (ds) RNA-binding domain (dsRBD) and tandem RNase III domains (RIIID) which dimerize to form a processing centre: RIIDa cuts the 3' strand and RIIDb cuts the 5' strand, generating a 3' two-nucleotide overhang. DGCR8 has two dsRBDs and functions as a molecular ruler to establish the precise Drosha cleavage site: approximately 11 bp from the stem-ssRNA junction (Han et al., 2006). pri-miRNA cleavage by Drosha is of critical importance because it determines one end of the mature miRNA. SNPs within these regions have been reported to block pri-miRNA processing to pre-miRNA, and consequently affect miRNA-mediated translational repression (Duan et al., 2007; Sun et al., 2009).

pre-miRNAs with a protruding 3' hydroxyl end and a 5'-phosphate end characteristic of RNase III processing (Lee et al., 2003) are exported to the cytoplasm via Exportin 5 (EXP5; Lund et al., 2004; Yi et al., 2003), in a Ran-GTP-dependent process (Bohnsack et al., 2004).

In mammals and flies, the 5' phosphorylated end generated by Drosha is recognized by Dicer's 5' pocket within the Piwi/Argonaute/Zwille (PAZ) domain. Since the RIIDs are positioned at the opposite end, a molecular ruler is set by the distance between the PAZ domain and the RIIDs. This distance defines the enzymatic cut approximately 22 bp from the 5' end – 5' counting rule (Park et al., 2011). As a result, the terminal stem bp and loop of the pre-miRNA are cleaved by the cytoplasmic RNase III Dicer, liberating a ~22 bp dsRNA (Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Knight and Bass, 2001).

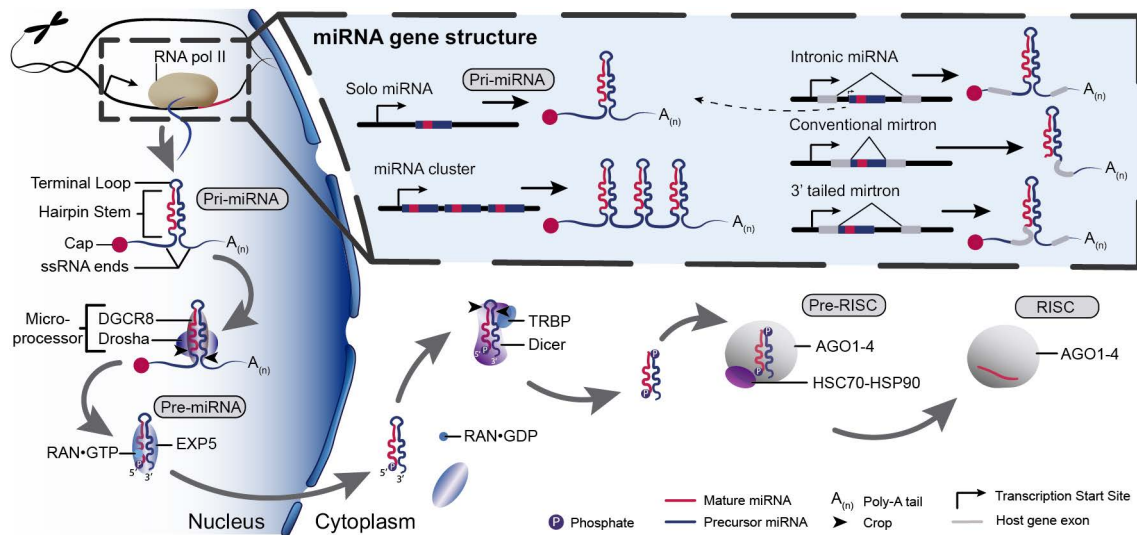
Dicer is often found in complex with dsRNA-binding proteins which have three dsRBDs, such as the Tar RNA binding protein (TRBP) in humans

(Chendrimada et al., 2005; Haase et al., 2005). TRBP appears to be important in the modulation of processing efficiency of some pre-miRNA (Fareh et al., 2016), generation of miRNA isoforms – isomiRs – (Lee and Doudna, 2012) and Argonaut (Ago) recruitment (Chendrimada et al., 2005).

After Dicer's cleavage, a ~22 bp miRNA duplex is formed and further incorporated by Ago (Hutvagner and Zamore, 2002; Meister and Tuschli, 2004; Mourelatos et al., 2002). Ago unwinds the duplex and discards one of the strands, whilst maintaining the other to form the RNA-induced silencing complex (RISC). Ago loading of the dsRNA requires the heat shock cognate 71 kDa (HSC70) protein and the heat shock protein 90 (HSP90) chaperone machinery. These proteins drive the conformational opening of Ago, in an ATP-dependent manner (Iki et al., 2010; Iwasaki et al., 2010; Johnston et al., 2010; Miyoshi et al., 2010).

After the incorporation of the dsRNA, forming a pre-RISC, the miRNA duplex suffers unwinding and the two strands are separated, resulting in the retention of the anchored strand – guide strand (“-5p” suffix) – and disposal of the other – passenger strand (“-3p” suffix). Strand selection is biased towards weaker hydrogen bonding at the 5' end (Khvorova et al., 2003; Schwarz et al., 2003). Weaker hydrogen bonding means that the structure is more flexible and therefore more easily disrupted by Ago. In fact, human miRNAs are biased towards having an A or U as their 5' terminal nucleotide (Baek et al., 2008; Grimson et al., 2007; Nielsen et al., 2007; Schirle et al., 2014). Additional destabilising elements come in the form of bulges, mismatches and gaps within the hairpin stem which increase the unwinding efficiency. Passenger strands are also incorporated into Ago, albeit in much lower frequency, and are able to exert regulatory functions (Yang et al., 2011).

After sorting, RISC is formed and the mature miRNA guides Ago by binding mRNA complementary sequences.



**Figure 1.5 miRNA gene types and canonical biogenesis.** miRNA genes can be found in several genomic contexts, from intergenic to intragenic localizations. If they are in close proximity to each other, forming a cluster, they can constitute a polycistronic transcription unit, generating a single primary miRNA (pri-miRNA). If they are located in intronic regions of protein-coding genes, they can, or not, share the promoter of the host gene, generating different pri-miRNAs. Moreover, short introns can give rise to pri-miRNAs which lack the lower stem (which mediates recognition and cleavage by the micro-processor). These miRNAs are thus processed differently – non-canonical biogenesis – than solo miRNAs. miRNAs are largely transcribed by RNA polymerase II (RNA pol II) into a pri-miRNA which is recognized and cleaved by the micro-processor. Cleavage by Drosha generates a precursor miRNA (pre-miRNA), which is exported to the cytoplasm in a Ran-GTP-dependent manner. Once in the cytoplasm, the pre-miRNA is further cleaved by Dicer, generating a RNA duplex. Dicer is often found in complex with TRBP. Following Dicer processing, the RNA duplex is loaded into the human ArgonAUT 1-4 (Ago 1-4) with the help of HSC70 and HSP90, which hydrolase ATP (not shown). The passenger strand is discarded, whereas the guide strand remains in the Ago protein, forming the RNA-induced silencing complex (RISC). single-stranded RNA (ssRNA), Di George critical region 8 (DGCR8), Guanosine triphosphate (GTP), Exportin 5 (EXP5), Tar RNA binding protein (TRBP), Heat shock cognate 71 kDa protein (HSC70), Heat shock protein 90 (HSP90).

### 1.4.2 miRNA targeting

Contrary to plant's miRNA, which have almost perfect Watson-Crick pairing with its target and direct RNA interference (RNAi) -like mRNA cleavage (Jones-Rhoades et al., 2006; Rhoades et al., 2002), metazoan miRNA commonly bind imperfectly, inducing translational repression and mRNA deadenylation and decay.

Pairing to the 5' segment of the mature miRNA, the so called “seed” (nucleotides 2-7) region, is known to be of most significance (Lewis et al., 2003). Seed nucleotides are the first nucleotides exposed by Ago during targeting (Schirle et al., 2014). As a result, contiguous and perfect Watson-Crick

base pairing to the seed – canonical binding – is considered the most stringent requirement for miRNA binding in most target-prediction algorithms.

miRNA binding can also include centered pairing where there is contiguous, about 11-12 nucleotides long, Watson-Crick pairing in the center of the miRNA (Shin et al., 2010). Also, very extensive pairing of the miRNA to the mRNA target can occur and induce RNAi-like cleavage instead of translational repression (Davis et al., 2005; Karginov et al., 2010; Shin et al., 2010; Yekta et al., 2004). However, these non-canonical sites are very rare as compared to canonical sites. Recently, they were shown to not mediate target repression, despite binding the miRNA (Agarwal et al., 2015). **Figure 1.6** represents the several types of miRNA binding.

**Figure 1.6 Canonical and non-canonical miRNA binding.** (Adapted from TargetScan website).

instance, the number of binding sites within the mRNA can affect repression since each additional site can cumulatively contribute to it (Grimson et al., 2007; Nielsen et al., 2007). Conversely, fewer miRNA target-sites across the transcriptome tend to increase silencing efficacy (Arvey et al., 2010; Nielsen et al., 2007).

Also, the context of the binding site can explain why a particular site is more effective in one place than it is in another. Effective sites tend to reside outside the ribosome path (first 15 nucleotides of the 3'UTR; Grimson et al., 2007). Additionally, 3'UTRs with high local (Grimson et al., 2007; Nielsen et al., 2007) and total (Hausser et al., 2009; Robins and Press, 2005) AU content tend to be more accessible to the silencing complex. Shorter distances between the site and the 3'UTR terminus (Gaidatzis et al., 2007; Grimson et al., 2007; Majoros et al., 2007) as well as shorter 3'UTR length (Betel et al., 2010; Hausser et al., 2009; Reczko et al., 2012; Wen et al., 2011) also tends to boost silencing efficacy by increasing RISC accessibility. Similarly, lower mRNA stability (Ameres et al., 2007; Kertesz et al., 2007; Long et al., 2007; Robins et al., 2005; Tafer et al., 2008) and higher seed pairing stability (Garcia et al., 2011) impart higher silencing efficacy.

### 1.4.3 miRNAs in common complex diseases

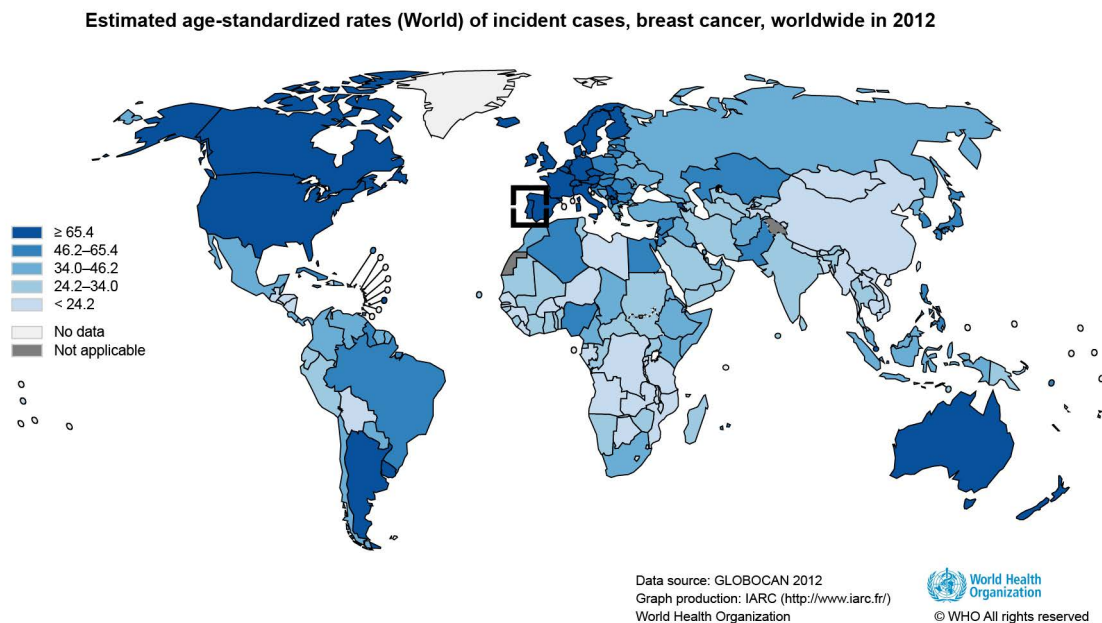
miRNAs have been demonstrated to play major roles in a wide range of human diseases (Ardekani and Naeini, 2010; Giza et al., 2014; Li and Kowdley, 2012), including cancer (Ha, 2011; Iorio et al., 2005; Yan et al., 2008; Zhong et al., 2012). Many miRNAs are de-regulated in primary human tumours (Lu et al., 2005) and even some miRNAs have been termed as “oncomirs”, for their tumour promoting role (Jiang et al., 2010; 2007; Yang et al., 2015), and/or tumour suppressors (Wang et al., 2016; Wu et al., 2009; Xu et al., 2017; Zhao et al., 2016).

There is strong, albeit episodic, evidence of genetic variation modulating the risk to complex diseases by altering miRNA regulation (Hu et al., 2016; Landi et al., 2008), including in BC (Brewster et al., 2012; Nicoloso et al., 2010).

## 1.5 BREAST CANCER

### 1.5.1 Epidemiology

Besides non-melanoma skin cancer, BC is the most common malignancy among women world-wide and of special emphasis in developed countries (Ferlay et al., 2013) – as accentuated by **Figure 1.7**. Just in Portugal, over 6000 new cases arise every year (Ferlay et al., 2013), a number which is still expected to increase, as epidemiological studies have undoubtedly showed us that BC is a disease of affluent societies with an acquired “Western-lifestyle”, marked by the combination of low physical activity and a highly caloric animal-based diet.



**Figure 1.7 Globocan 2012 illustrating the world’s breast cancer burden.** Breast cancer incidence is highest amongst developed countries, including Portugal (highlighted). Incidence rates are expressed per 100 000 person-years (Adapted from Ervik et al., 2016).

### 1.5.2 Aetiology

BC is a complex, multifactorial disease, combining genetic and environmental risk factors (Collaborative Group on Hormonal Factors in Breast Cancer, 2001). Gender and age are the most highly associated risk factors with BC development; and most cases are found in women aged 50 years or older.

Furthermore, a woman's reproductive lifestyle also confers increased risk: **(i)** being menstruated before age 12 and/or **(ii)** starting menopause after age 55 augment the exposure to oestrogen, a hormone whose role in BC aetiology is well established (Key and Verkasalo, 1999; Russo and Russo, 2006; Travis and Key, 2003); and **(iii)** nulliparity or low parity, especially after age 35, also increases risk (Key et al., 2001).

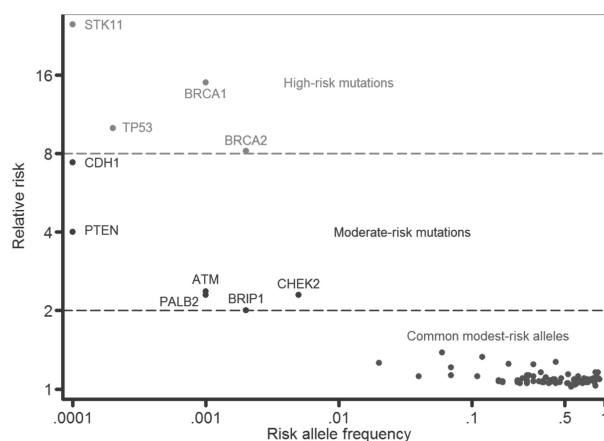
Several other environmental factors seem to contribute, although with lower impact, to BC risk such as radiation, alcohol, obesity and exogenous hormone intake. Higher impact of the latter environmental factors seems only to be conferred to post-menopausal women, as ovarian oestrogen and progesterone synthesis ceases and androgen production lowers.

As most common human neoplasms, **BC shows familial clustering** – the occurrence of the disease within some families more than what would be expected from the occurrence in the population. Mapping the genes responsible for inherited BC may allow the stratification of the population according to individual risk and improve screening measures that will contribute to identify early lesions, critical for the development of BC in the general population – sporadic BC.

### 1.5.3 Breast Cancer Familial Risk

Having a first-degree family member with BC, increases a woman's personal risk of developing it herself: in fact, this risk nearly increases an additional fold proportionally to the number of affected family members (Collaborative Group on Hormonal Factors in Breast Cancer, 2001), indicating a role for genetics in BC aetiology. However, familial clustering is a mere epidemiologic pattern and does not reveal the causes of that phenotypic aggregation and, besides genetic predisposition, exposure to certain environmental factors by some family members can mimic genetic inheritance patterns. Evaluation of the prevalence of cases in twins, isolates the genetic effect by eliminating the presumably shared equal environment. As such, the observed higher concordance of BC cases in monozygotic twins compared to dizygotic twins (Ahlbom et al., 1997; Mack et al., 2002) further corroborates the genetic component of BC risk.

It is estimated that 5-10% of BCs are attributable to familial inheritance (Claus et al., 1991; Newman et al., 1988; Rowell et al., 1994). Due to the high incidence of BC, the inherited form of BC is actually a common genetic disease, despite being a small fraction of the BC burden (King et al., 1993). In the last 30 years, molecular genetic studies have identified several risk *loci* responsible for BC familial risk, some of which are illustrated in **Figure 1.8**.



**Figure 1.8 Breast cancer susceptibility *loci* by risk allele frequency and conferred relative risk.** (Adapted from Ghoussaini et al., 2013)



#### 1.5.3.1 High-risk mutations

High-risk mutations were the first identified risk *loci* in BC, through linkage studies in high-risk families (Hall et al., 1990; Narod et al., 1991; Wooster et al., 1994, 1995). A high proportion of individuals carrying these risk *loci* develop the disease (high penetrance). As a result, they are more easily identifiable due to near-Mendelian inheritance patterns, where genotype is highly associated with phenotype. Also, they are very rare in the population and therefore only explain 20% of the family cases (Easton, 1999).

The most famous example of high-risk *loci* in BC is of mutations in the breast cancer 1/2 (*BRCA1/2*) genes. These were identified through linkage analysis in families with frequent and early-onset of the disease and who had also ovarian cancer cases (Hall et al., 1990; Narod et al., 1991; Wooster et al., 1994, 1995). Linkage analysis is based on the observation that closely located genes on a chromosome remain linked during meiosis, i.e., they are inherited together. Therefore, by using a genetic marker or a DNA segment with known physical location in a chromosome and tracking its inheritance, unknown disease-related genes can eventually be identified if the genetic markers are co-inherited with the trait of interest.

Other examples of high-prevalence BC risk *loci* include mutations in the tumour protein 53 (*TP53*) gene predisposing to the Li-Fraumeni Syndrome, where risk of BC is the highest among its cancer spectrum (Malkin et al., 1990); and in the serine/threonine kinase 11 (*STK11*) gene accountable for the Peutz-Jeghers syndrome (Chen and Lindblom, 2001; Hemminki et al., 1998).

Together, these BC risk *loci* only accounted for a small fraction of the familial risk and subsequent linkage studies failed to identify BC genes that conferred moderate risk. This suggested that: **(i)** there were more risk *loci* to be held liable; and **(ii)** BC was largely polygenic, having a large number of *loci* contributing to increased susceptibility, each conferring its own small effect.

### 1.5.3.2 Moderate-risk mutations

After the identification of *BRCA1/2* and their later implication in DNA repair pathways (Connor et al., 1997; Patel et al., 1998; Sharan et al., 1997; Xu et al., 1999), it was proposed that similar, undiscovered, genes that either **(i)** interacted with them or **(ii)** shared the same pathways, could be key players in BC development. This led to the employment of mutational screening studies of candidate genes in large case-control cohorts. Examples of these screen-identified risk *loci* in BC include mutations in the checkpoint kinase 2 (*CHEK2*; Bell, 1999; Meijers-Heijboer et al., 2002), Ataxia-Telangiectasia mutated (*ATM*; Easton, 1994; Renwick et al., 2006; Swift et al., 1976; Thompson et al., 2005), BRCA1-interacting protein 1 (*BRIP1*; Cantor et al., 2001; Seal et al., 2006), and partner and localizer of BRCA2 (*PALB2*; Rahman et al., 2007; Xia et al., 2006) genes. These moderate penetrance risk *loci*, conferring 2-3 fold increased risk, account for 25% of the familial risk (Easton, 1999; Ghoussaini et al., 2013), still leaving room for over 50% of BC familial cases unexplained – missing heritability.

### 1.5.3.3 Common modest-risk alleles

Additional identification of 16% of the BC familial risk (Michailidou et al., 2015) was achieved via association studies throughout the whole genome, where chromosomal regions were associated to BC in large-scale case-control cohorts. These studies, termed GWAS, thoroughly incremented the statistical power, which allowed increased resolution for the unbiased identification of common variants (minor allele frequency, MAF, >5%) that were associated with small cumulative risks (approximately 1.5 fold).

To date, over 100 *loci* were associated with BC through GWAS, consistent with the polygenic model. However, there are still approximately half of the BC familial risk unexplained.

#### 1.5.3.3.1 GWAS

GWAS are association studies in large-scale case-control cohorts, in which SNPs usually with a MAF of at least 5% are genotyped and then their frequencies are compared amongst groups. If the allele frequencies between the disease and control group are significantly different to what would be expected by chance, then they are considered to be associated with the disease.

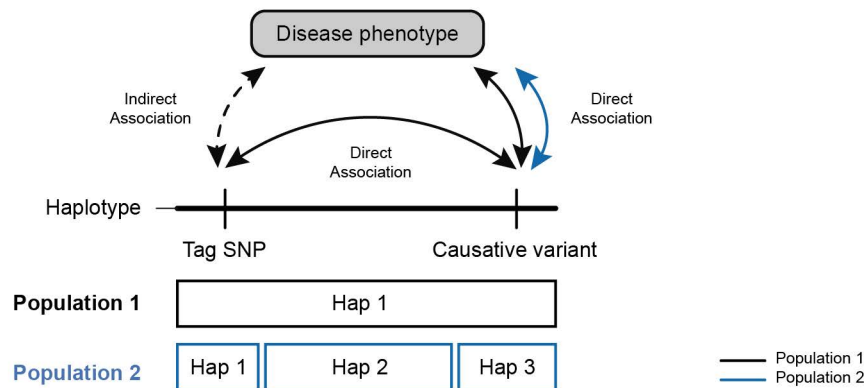
These studies are based on the observation that some SNPs have a non-random genetic association. When the alleles at two *loci* are not independent, i.e., the haplotype frequencies are different from what is expected at random association (equilibrium), it is said they are in linkage disequilibrium (LD). As recombination events occur in a population, from generation to generation, contiguous chromosomal stretches in linkage are broken apart until, eventually, they achieve linkage equilibrium and the haplotype frequencies are independent.

LD is generally reported in terms of the square of the correlation coefficient between two indicator variables – one representing the presence or absence of a particular allele at the first *locus* and the other representing the presence or absence of a particular allele at the second *locus* (Rosenberg and VanLiere, 2009; VanLiere and Rosenberg, 2008) –,  $r^2$ , which ranges from 0 to 1. Having an  $r^2=1$  for two SNPs indicates that they convey equivalent information (one allele of the first SNP is always observed in the presence of one allele of the second SNP) and, as a result, there is no need to genotype both SNPs because only one is sufficient to capture the allelic variation – tag SNPs. **Figure 1.9** illustrates the haplotype blocks in different populations and the relation of the haplotype tagging SNP, genotyped in GWAS, with the true causal variant.

Given that the LD signal decays substantially with distance between alleles, it was initially estimated that half a million SNP *loci* would be needed to cover the entire human genome of non-African descendants. Moreover, the 1000 Genomes Project has now identified several tens of million SNPs through whole-genome sequencing technologies.

Thanks to the Haplotype Map (HapMap) Project, which by the beginning of 2005 had genotyped over 1.1 million SNPs across four populations (Thorisson et al., 2005), and to the recent advances in microarray technology, which allowed the systematic genotyping of large amounts of SNPs across the genome, the first successful GWAS arrived later that year (Klein et al., 2005).

GWAS studies, taking advantage of this knowledge, could then be performed by genotyping a much smaller set of SNPs, the tag SNPs, which could report on 100's of other SNPs in high LD within the same haplotypes. Nevertheless, if this simplified and cut the cost of genotyping, on the other hand, generated a new problem. When a SNP is associated with risk, the true causal variant is not necessarily the tag SNP being genotyped, and so further fine-mapping and educated prioritisation needs to be carried to identify this causative variant.



**Figure 1.9 Schematic diagram of the haplotype blocks in two populations and relation to GWAS.** In population 1, the GWAS genotyped SNP, which is indirectly associated with the disease, is in the same haplotype block as the causative variant. Although the direct association between the unstudied causal variant both the genotyped SNP and the disease phenotype cannot be observed, if the variants have a high  $r^2$  between them, we might be able to detect the indirect association between the genotyped SNP and disease phenotype. However, population 2 has suffered a recombination event and haplotype 1 is spliced into three haplotypes. Here, the genotyped SNP and the causative variant are no longer in the same haplotype block nor in high linkage disequilibrium (LD) with each other, meaning that if we were to genotype both SNPs in population 2, only the causative SNP would remain associated with the disease. This illustrates the use of fine-mapping studies to refine the association signal in GWAS (Adapted from Balding and Balding, 2006; Ghousaini et al., 2013). Genome-wide Association Studies (GWAS), Single Nucleotide Polymorphism (SNP), Haplotype (Hap).

Since the first published study, tens of thousands of SNP *loci* have been associated with different traits (MacArthur et al., 2017). The first BC GWAS arrived ten years ago (Easton et al., 2007) and was the first one carried for a common cancer.

## 1.6 PRIORITIZING C/S-REGULATORY CANDIDATES

GWASes have provided insightful information about the importance of normal genetic variation in common disease. They have identified many common variants, often in non-coding regions, to confer modest risk. However, the vast majority have no established biological relevance. Since these studies provide the statistical association of a *locus* to a complex trait/disease but not its cause, the output of identified risk SNPs is a representation of the haplotype block, where correlated SNPs are statistically indistinguishable. Fine-mapping studies (the process of refining the association signal) together with functional analysis are required to fully identify the causal variant.

For the few existing functional studies, these variants were shown to *cis*-regulate gene expression levels of both close and distant target genes (Meyer et al., 2008; Dunning et al., 2009, 2016; Maia et al., 2012;). In the first functional study reported, the most significant hit of the first BC GWAS was analysed (Meyer et al., 2008). They found two SNPs overlapping two intronic enhancers, which modified the binding of the transcription factors C/EBP $\beta$ , and the pair OCT-1 and Runx2. The authors showed that by altering the affinity of the TF binding, the levels of expression of the gene *FGFR2* (fibroblast growth factor receptor 2) were modified, with the risk haplotype identified in the GWAS being associated with a small elevation in this expression (Meyer et al., 2008).

Further studies mainly focused on TF binding prediction, *DNase*I hypersensitivity (indicative of open chromatin, possibly indicating regions of TF binding) and chromatin conformation capture data (indicative of physical contact between candidate enhancers at promoters of target genes), therefore overlooking other mechanisms capable of generating allelic imbalances in gene expression (Fachal and Dunning, 2015).

As explained before, miRNA binding and biogenesis are both sequence dependent, and therefore susceptible to be altered by genetic variation across the genome. The same is true for splicing, in which specific sequences determine the binding site of splice factors.

As a result, **we hypothesized that miRNA-mediated *cis*-regulation may also be important to BC risk.** *Cis*-acting variants, associated with risk to BC, may affect this mechanism via the structure and stability of the miRNA itself (is located within the miRNA gene sequence), or affect the miRNA::mRNA binding affinity (if located both in the seed region of the miRNA, or the target site in the regulated mRNA). In either way, the levels of expression of a target gene could be altered and increase cancer susceptibility.

## 1.7 PRELIMINARY DATA

As mentioned above, the most robust approach to identify the effect of *cis*-acting genetic variation is the analysis of DAE, and so DAE data can be used to prioritize GWAS-identified SNPs for their greater regulatory potential.

Previously, in Prof. Ana Teresa Maia's group, SNP microarray-derived DAE genomic mapping was performed using 64 samples of normal breast tissue (Xavier et al., 2016). From these experiments, Maia and colleagues found that approximately 20% of genes expressed in breast were controlled by *cis*-regulatory variants.

DAE mapping is a useful tool for searching *cis*-regulatory variants which affect the expression of BC-associated risk genes. As a result, our DAE genomic mapping was previously overlapped with BC GWAS published results in order to prioritize candidate risk *loci* for functional validation. This resulted in the identification of 200 candidate *loci* which showed evidence of being *cis*-regulated and were associated with BC risk. From those, 15 *loci* displaying GWAS associated SNPs in strong LD with SNPs displaying DAE (Xavier et al., 2016). One of those *loci* is located in 5q14.2 and spanned across three genes (*ATG10* – autophagy related 10, *RPS23* – ribosomal protein 23, and *ATP6AP1L* – ATPase, H<sup>+</sup> transporting, lysosomal accessory protein 1-like).

## 2. AIM

Here we aim to determine the possible contribution of *cis*-regulatory SNPs influencing miRNA post-transcriptional regulation to BC susceptibility.

We will evaluate the possible miRNA-mediated mechanisms that lie at the base of specific gene expression variation with the help of several biomedical and bioinformatic tools and databases, and perform *in vitro* validation of the best candidates identified.

More specifically, we will identify candidate risk-associated *loci* that affect:

- The miRNA gene sequence, and possibly its hairpin stability or seed region;
- The miRNA recognition site on target genes identified via GWAS.





## **CHAPTER III**

# **Materials and Methods**



### 3. MATERIALS AND METHODS

#### 3.1 DATASET

GWAS-significant SNPs for BC susceptibility were retrieved from the NHGRI-EBI Catalogue of published genome-wide association studies (MacArthur et al., 2017), available at [www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas) (Accessed 13/02/2017), using a p-value cut-off of  $\leq 5 \times 10^{-6}$ . This list was further increased by 15 GWAS-significant SNPs identified by Michailidou, K. and colleagues (Michailidou et al., 2015) plus 79 *loci* provided in the same study.

#### 3.2 miRNA-TARGET PREDICTION

##### 3.2.1 R programming language and RStudio® IDE

To implement a systematic, automated and reproducible *in silico* analysis aiming to determine the possible contribution of *cis*-regulatory SNPs in miRNA post-transcriptional regulation, we made use of the R programming language (R Core Team, 2017). R is a free software environment for statistical computing and graphics that can be easily extended via packages, mostly open source, several of which have already been successfully implemented to analyse biological data. Additionally, its core can be used to implement loops, branches and modular programming through functions that are computer-interpreted to perform specific instructions.

Accordingly, **a portion of the time dedicated to the present dissertation was spent learning R programming and package implementation.**

Programming in R (version 3.3.2) was done in RStudio® (version 0.99.903), an integrated development environment (IDE) for R, and ran under Ubuntu (release 16.04.1 LTS) in a 64-bit Linux/GNU platform.

### 3.2.2 SNP query with SNAP (Broad Institute)

GWASes use a tag-SNP approach to the identification of risk loci, meaning that any SNP in tight LD with the GWAS-SNP can be the true causal variant. Therefore, a list of proxy SNPs, those in high LD with a tag-SNP, were identified to be added to the analysis.

Proxy SNPs were generated using the SNP Annotation and Proxy Search (SNAP) online tool (version 2.2; Johnson et al., 2008), available at <http://archive.broadinstitute.org/mpg/snap/ldsearch.php>, using genotype data from the 1000 Genomes Pilot 1 for the CEU population (Utah Residents with Northern and Western European Ancestry). SNAP is a web-based service for the rapid retrieval of LD proxy SNPs from a given input SNP query, based on  $r^2$  thresholds and/or physical distance. Settings were established as: distance limit of 500 kb, and an  $r^2$  threshold of  $\geq 0.8$ .

SNAP output was imported into the RStudio® IDE and a vector of unique GWAS-significant SNPs and their proxies was obtained. Additional relevant data, including their genomic localization as annotated in release 87 of the Ensembl database (McLaren et al., 2016), was retrieved via the *getBM* function from the biomaRt R package (version 2.30.0; Durinck et al., 2005, 2009). Ensembl is a genome browser for vertebrate genomes that supports comprehensive research in several areas, including sequence variation (Yates et al., 2016). One of Ensembl's many tools is the data-mining tool BioMart, which allows to export custom sets of data from the Ensembl public data repository. biomaRt provides an R-based interface to databases implementing the BioMart software suite, thus enabling the retrieval of large amounts of data in an uniform way.

A filter was then set for SNPs located inside PCGs (either 5'UTR, CDS or 3'UTR) or near/inside miRNA genes.

### 3.2.3 miRNA-target binding prediction algorithms

Following the initial SNP query, two distinct miRNA-target binding prediction algorithms were employed: TargetScan v7.1 (Agarwal et al., 2015) and miRanda v3.3a (Enright et al., 2003a). These were selected amongst other prediction algorithms for their methodology and viable implementation in R.

As none of the algorithms were prepared to process SNP allele queries, we developed a logical approach to overcome this issue: **(i)** first each SNP allele located in PCGs was independently evaluated for putative miRNA-binding through the algorithm; **(ii)** allele-specific miRNA binding predictions for each SNP was obtained by calculating the difference between the outputs of the corresponding SNP alleles. The analysis flowchart is presented in **Figure 3.1**.

#### 3.2.3.1 TargetScan

##### 3.2.3.1.1 Tool Description

TargetScan predicts conserved and non-conserved targets of miRNA and outputs a context++ score for canonical miRNA binding sites. This score was developed by applying multiple linear regression models to each site type (6mer, 7mer-A1, 7mer-m8, 8mer) using 14 features described in **Table 3.1**. The lower the context++ score, the higher the probability of targeting/repression (Agarwal et al., 2015).

##### 3.2.3.1.2 miRNA::mRNA interaction predictions

For all SNPs located at the 3'UTR of PCGs, each allele was analysed for miRNA::mRNA interactions using the default settings of the predictive algorithm TargetScan v7.1 (Agarwal et al., 2015). Source code and data, including target and miRNA mature sequences, were obtained from the TargetScan website (available at [http://www.targetscan.org/cgi-bin/targetscan/data\\_download.vert71.cgi](http://www.targetscan.org/cgi-bin/targetscan/data_download.vert71.cgi)).

**Table 3.1 Features included in the context++ score calculated by TargetScan.** (Adapted from Agarwal et al., 2015) Untranslated region (UTR), small RNA (sRNA), open reading frame (ORF).

Feature	Description
<b>miRNA</b>	
3'UTR target-site abundance	Number of sites in all annotated 3'UTR
Predicted seed-pairing stability	Predicted thermodynamic stability of seed pairing
sRNA position 1	Identity of nucleotide at position 1 of the sRNA
sRNA position 8	Identity of nucleotide at position 8 of the sRNA
<b>Site</b>	
Site position 8	Identity of nucleotide at position 8 of the site
Local AU content	AU content near the site
3' supplementary pairing	Supplementary pairing at the miRNA 3' end
Predicted structural accessibility	log10(Probability that a 14-nucleotide segment centred on the match to sRNA positions 7 and 8 is unpaired)
Minimum distance	log10(Minimum distance of site from stop codon or polyadenylation site)
Probability of conserved targeting	Probability of site conservation, controlling for dinucleotide evolution and site context
<b>mRNA</b>	
ORF length	log10(Length of the ORF)
3'UTR length	log10(Length of the 3'UTR)
3'-UTR offset-6mer sites	Number of offset-6mer sites in the 3' UTR
ORF 8mer sites	Number of 8mer sites in the ORF

3'UTR multiple sequence alignments across 84 vertebrate species obtained from the TargetScan website were used as reference. For each SNP located in a specific human 3'UTR sequence alignment, independent text files containing either the common or the variant allele in the human sequence were generated from the reference. To each file, the correspondent 3'UTR sequence alignments for the remaining 83 species in the reference dataset were added so that target conservation could be taken into account by the algorithm. Default instructions available for context++ score calculation were followed for each SNP allele and differences between each correspondent output file were obtained.

The R markdown containing the complete pipeline and description of the R script created for the TargetScan-based analysis can be found in **Annex D**.

### 3.2.3.2 miRanda

#### 3.2.3.2.1 Tool Description

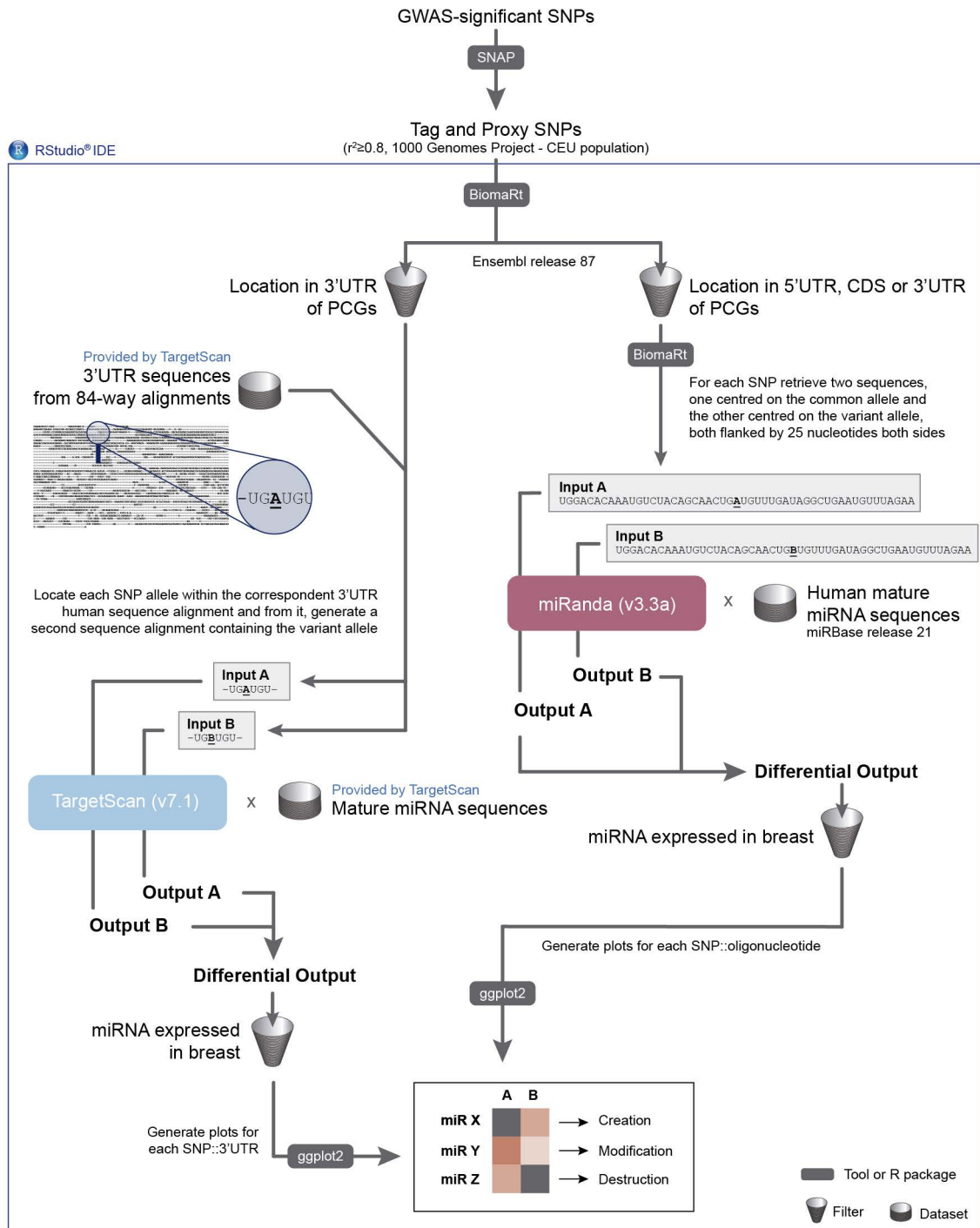
miRanda is an algorithm that detects potential miRNA target sites in genomic sequences (Enright et al., 2003b). First, a local alignment is carried out between the miRNA sequence and the mRNA sequence. The alignment is scored based on sequence complementarity. The G:U wobble base pair is also permitted, although it scores less than the Watson-Crick matches. Next, RNA duplex thermodynamic stability is estimated based on the RNAlib library (part of the ViennaRNA package) and the minimum free energy (kcal/mol) is calculated. RNAlib is a library for folding and comparing RNA secondary structures.

#### 3.2.3.2.2 miRNA::mRNA interaction predictions

For SNPs mapped to PCG's (either annotated as 5'UTR, CDS or 3'UTR variants), the miRanda algorithm v3.3a (Enright et al., 2003a) was employed using a cut-off score  $\geq 80$  and a minimum free energy  $\leq -16$  kcal/mol (as previously described by Nicoloso et al., 2010) for a retrieved target-sequence centred on each allele and flanked by 25 nucleotides both sides (Ensembl release 87 via biomaRt version 2.30.0).

Human miRNA mature sequences were subsetted from the *mature.fa* document (available at: <ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz>) retrieved from miRbase (release 21), an online searchable database of published miRNA sequences and annotations. miRanda source code was obtained from the microRNA.org website (available at: [http://cbio.mskcc.org/microrna\\_data/miRanda-aug2010.tar.gz](http://cbio.mskcc.org/microrna_data/miRanda-aug2010.tar.gz)).

Differences between each correspondent output file for a specific SNP-oligonucleotide were generated through the development of a parser.



**Figure 3.1 Flowchart of differential miRNA-target prediction analysis.** GWAS-significant SNPs (tag) and their proxies in high linkage disequilibrium ( $r^2 \geq 0.8$ ), retrieved from the SNP Annotation and Proxy Search (SNAP) online tool, were imported to the RStudio® integrated development environment (IDE). Through the biomaRt R package, relevant data related to the SNPs was retrieved from the Ensembl database. SNPs were filtered by their location in protein-coding genes (PCGs). If located in the 3' untranslated region (UTR) of PCGs, then both the TargetScan and miRanda algorithm would be applied. If located in the 5'UTR or in the coding sequence (CDS) of PCG's, then only the miRanda algorithm would be applied. The algorithms have distinct inputs. Continued on next page.



**Figure 3.1** Continued. In the case of the TargetScan-based analysis, the input target sequence corresponds to the full human 3'UTR sequence aligned across 84 vertebrate species (also present in the input; not shown). The 84-way multiple sequence alignments were retrieved from the TargetScan website (available at: [http://www.targetscan.org/vert\\_71/vert\\_71\\_data\\_download/UTR\\_Sequences.txt.zip](http://www.targetscan.org/vert_71/vert_71_data_download/UTR_Sequences.txt.zip)) and set as reference. 3'UTR sequence alignments containing the variant SNP allele in the human sequence were generated from the latter. As for the miRanda analysis, 51-nucleotide target sequences containing either the common allele or the variant allele were retrieved from the Ensembl database via biomaRt. Human miRNA mature sequences were retrieved from miRbase. In both analysis, input A (sequence containing the common allele A) and input B (sequence containing the variant allele B) were ran through the algorithms independently. Their correspondent output file (A and B) was later compared and differences between them were extracted – differential output. The latter was then filtered for miRNAs expressed in normal or tumoral breast tissue. Axial plots illustrating the output of miRNA allele-specific binding were generated using the ggplot2 R package.

### 3.2.4 miRNA expression in breast

miRNAs exhibit tissue-specific expression. Therefore, a final subsetting step for miRNAs expressed in normal or tumoral breast tissue was applied to each differential output file.

Normal breast miRNA expression data was obtained from the miRmine database (Panwar et al., 2017) (Experiment Accession No.: SRX513286), a database of human miRNA expression profiles. For simplicity's sake, miRNAs with expression values >0 reads per mapped million (RPM) were considered as expressed.

miRNAs expressed in breast cancer were obtained from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) miRNA landscape study (Dataset Accession No.: EGAD00010000438). miRNA names (miRbase v16) were converted to their more recent nomenclature (miRbase v21) through the *translateMiRNAName* and *MiRNANameConverter* functions from the *miRNANameConverter* R package (version 1.3.1; Haunsberger et al., 2016).

### 3.2.5 Plot generation

Axial plots illustrating the output of miRNA allele-specific binding were generated for each SNP using the ggplot2 R package (version 2.2.1). ggplot2 is a system for declaratively creating graphics (Wickham, 2010).

### 3.2.6 DAE SNP ranking

In order to prioritize putative *cis*-regulatory SNPs for functional validation, SNPs with predicted allele-specific miRNA binding were filtered based on previous evidence of DAE.

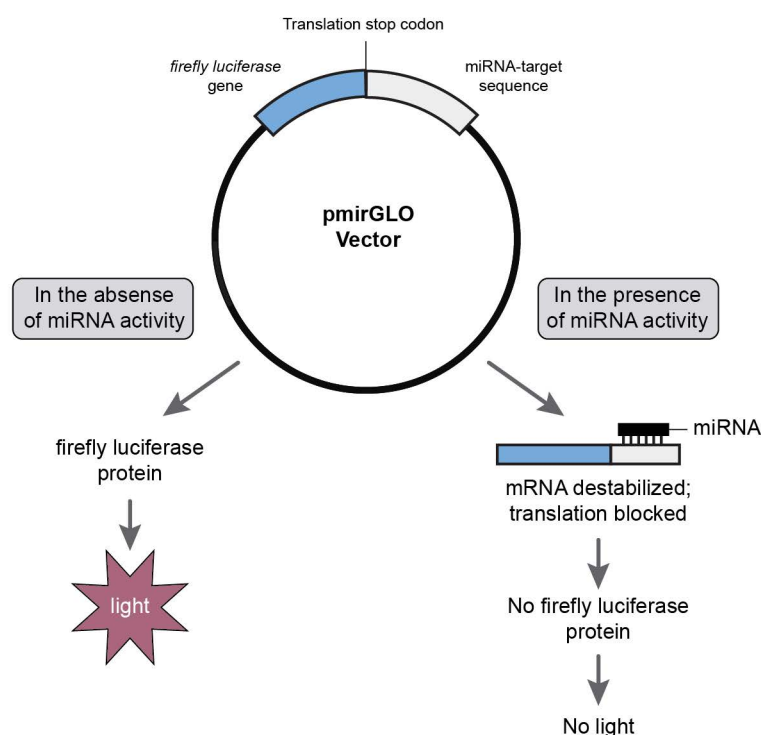
DAE genomic mapping was previously done in Ana-Teresa Maia's group using mRNA expression data from 64 normal breast tissue samples obtained by microarray technology (Xavier et al., 2016). Oligonucleotide probes in SNP microarrays distinguish the signal between both alleles of transcribed SNPs, enabling allele-specific measurements. When the allelic ratio in heterozygous individuals is different from 1, then there is DAE (Liu et al., 2012).

## 3.3 FUNCTIONAL VALIDATION

### 3.3.1 Luciferase Reporter Gene Assays

To evaluate the putative allele-specific miRNA binding of hsa-miR-21-3p to rs6884232, we carried a functional study using a dual-luciferase system.

The luciferase reporter assay is commonly used to assess transcriptional activity. Its wide use is due to its simplicity, sensitiveness and almost instantaneous quantitative measurements of gene expression. Luciferase is an oxidative enzyme found in several organisms (such as firefly and renilla), enabling them to emit light (luminescence). This system can be applied to evaluate the effect of miRNA-mediated post-transcriptional regulation of target genes, i.e. loss of transcriptional activity, as illustrated in **Figure 3.2**. This is achieved by cloning the predicted miRNA-target sequence at the 3'UTR of the *luciferase* gene.



**Figure 3.2 Mechanism of action of a luciferase reporter gene assay developed to test miRNA-mediated post-transcriptional regulation of target genes.** Exemplified is the pmirGLO vector (Promega®, Cat. #E1330, Part No. E133A). The putative miRNA-target sequence is located in the 3' untranslated region of the *firefly luciferase* gene. If a miRNA is not able to bind the putative miRNA target, there is no silencing effect and the firefly luciferase protein is produced. In the presence of firefly luciferase substrate (not shown), firefly luciferase is able to catalyse that substrate and generate luminescence (light). On the other hand, if a miRNA is able to bind the transcribed miRNA-target sequence, then the *firefly luciferase* gene is post-transcriptionally silenced and protein translation does not occur. As a result, in the presence of the luciferase substrate (not shown), there is no luminescence observed because there is no luciferase protein to catalyse that substrate. (Adapted from the pmirGLO Dual-Luciferase miRNA Target Expression Vector protocol)

### 3.3.1.1 Oligonucleotide design and plasmid selection

Two small oligonucleotide pairs (Invitrogen®) were designed centred on each allele of rs6884232 flanked by 25 nucleotides each side, correspondent to its normal genomic context in the 3'UTR of *ATG10* (NCBI Accession No. NM\_031482.4: 1093-1143). The oligonucleotide pairs contained the overhang sequences for the *XhoI* (Forward) and *XbaI* (Reverse) restriction enzymes, which were used for establishing the 5' to 3' cloning orientation into the multiple cloning site (MCS) of the pmirGLO Dual-Luciferase miRNA Target Expression Vector (Promega®, Cat. #E1330, Part No. E133A). **Table 3.2** contains the sequence of each oligonucleotide.

**Table 3.2 rs6884232 oligonucleotide pairs.** *Xho*I (Forward) and *Xba*I (Reverse) overhangs are underlined, whereas rs6884232 alleles are both bold and underlined.

Allele	Sequence	Strand
<b>A</b>	<u>TCGAT</u> GGGACACAAATGTCTACAGCAACTG <b><u>A</u></b> TGTTTGATAGGCTGAATGTTTAGAA	Forward
	<u>CTAG</u> TTCTAAACATTCAGCCTATCAAACA <b><u>T</u></b> CAGTTGCTGTAGACATTTGTGTCCA	Reverse
<b>G</b>	<u>TCGAT</u> GGGACACAAATGTCTACAGCAACTG <b><u>G</u></b> TGTTTGATAGGCTGAATGTTTAGAA	Forward
	<u>CTAG</u> TTCTAAACATTCAGCCTATCAAACA <b><u>C</u></b> CAGTTGCTGTAGACATTTGTGTCCA	Reverse

The selected vector backbone has several key features: **(i)** enables bacterial selection for vector amplification by conferring ampicillin resistance through  $\beta$ -lactamase gene expression (*Amp<sub>r</sub>*); **(ii)** quantitatively reports miRNA activity by inserting miRNA-target sites in the MCS located 3' of the firefly luciferase reporter gene (*luc2*) and 5' of the Simian Virus 40 (SV40) late poly(A) signal; **(iii)** has a non-viral universal promoter (Human phosphoglycerate kinase, PGK, promoter) driving *luc2*, providing low translational expression; and **(iv)** has a humanized Renilla luciferase-neomycin resistance cassette (*hRluc-neo*), enabling control reporter normalization of *luc2*. The complete set of features is illustrated in the pmirGLO vector map of **Figure 3.3**.

### 3.3.1.2 Annealing, digestion and ligation

To obtain ds inserts, reconstituted oligonucleotide pairs in nuclease-free water (1 $\mu$ g/ $\mu$ L) were annealed in Oligo Annealing Buffer (Promega®, Cat. #E1330, Part No. C838A). Briefly, 2  $\mu$ L of each oligonucleotide (forward and respective reverse) were combined with 46  $\mu$ L of Oligo Annealing Buffer and heated at 90°C for 3 minutes followed by a 15-minute water bath at 37°C.

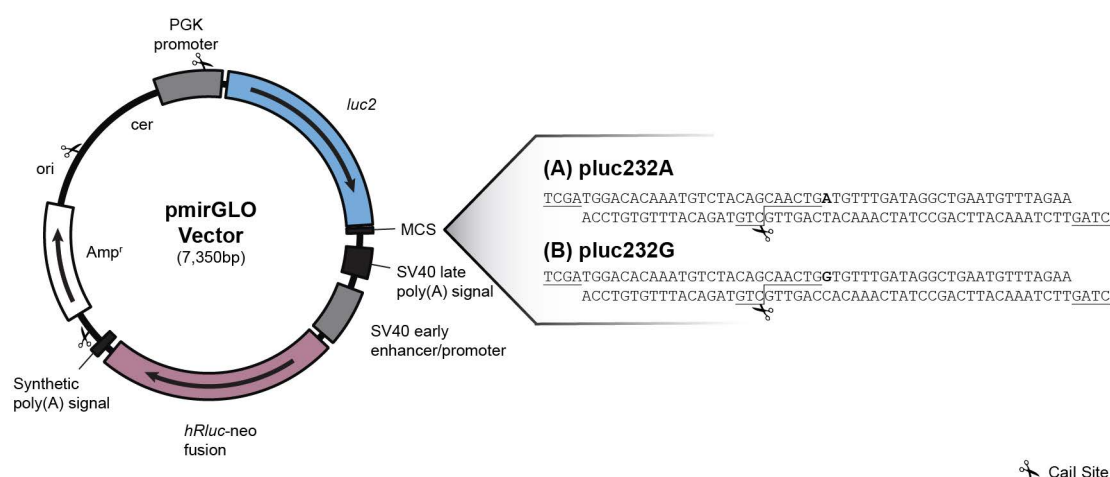
pmiRGLO vector was linearized with both the FastDigest *Xho*I (Thermo Scientific®, Cat. #FD0695) and FastDigest *Xba*I (Thermo Scientific®, Cat. #FD0685) restriction enzymes. Double digestion was performed according to manufacturer's protocol in FastDigest buffer. A 20-minute heat inactivation of the restriction enzymes was performed at 65°C.

Annealed oligonucleotides were ligated to the linearized vector using the T4 DNA Ligase (Thermo Scientific®, Cat. #EL0014). 2 ng of each annealed oligonucleotide were combined with 50 ng of linearized pmirGLO vector in a

sticky-end ligation protocol according to the manufacturer. Heat inactivation of the T4 DNA ligase was performed at 65°C for 10 min.

Both vector linearization and ligation were confirmed with a DNA 0.6% agarose gel electrophoresis in Tris-Borate-EDTA (TBE) buffer with 1 µg/µL of Green-safe. NZYDNA ladder III (NZYTech®, Cat. #MB04401) was used.

The pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 will be, from now on, referred as pluc232A, whereas the pmirGLO plasmid containing the G allele will be referred as pluc232G. **Figure 3.3** illustrates the composition of pluc232A and pluc232G.

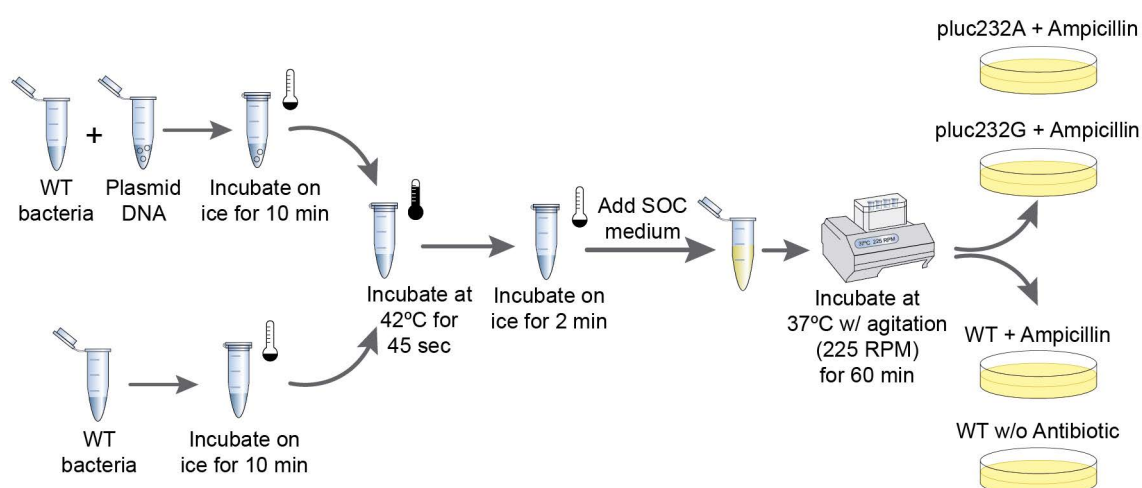


**Figure 3.3 pmirGLO vector and oligonucleotide inserts.** The pmirGLO vector features are illustrated. Oligonucleotide inserts were cloned into the MCS located immediately downstream of *luc2*. rs6884232 alleles are in bold, whereas the *Xho*I (forward) and *Xba*I (reverse) overhangs are underlined. The *Ca*I restriction site is illustrated and marked by a scissor, both in the oligonucleotide inserts and the vector backbone. Phosphoglycerate kinase (PGK), firefly luciferase reporter gene (*luc2*), Multiple Cloning Site (MCS), Simian Virus 40 (SV40), Humanized Renilla luciferase-neomycin resistance cassette (*hRluc-neo*), β-lactamase gene (*Amp<sup>r</sup>*), *ColE1*-derived plasmid origin of replication (*ori*), *ColE1*-derived recombination site (*cer*), pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G).

### 3.3.1.3 Bacterial transformation

To 100  $\mu$ L of JM109 competent *Escherichia coli* (Promega®, Cat. #L2001) we added 12 ng of plasmid DNA (either pluc232A or pluc232G). The solution was kept on ice for 10 minutes and incubated for 45 seconds at 42°C. 900  $\mu$ L of chilled super optimal broth with catabolite repression (SOC) medium was then added after a briefly 2 min incubation on ice. Bacterial recovery was done at 37°C with agitation (225 rpm) for 60 min. The same procedure was done for non-transformed *E. coli*, i.e., without plasmid DNA addition and, from now on, referred to as WT (as in “wild-type”).

In a 10-cm Petri dish with Luria-Bertani (LB) -agar (Lennox Formulation) we plated 100  $\mu$ L of the *E. coli* solution. Antibiotic (Ampicillin) was previously added to each LB-agar-containing Petri dish for a final concentration of 100  $\mu$ g/mL. To one of the Petri dishes, no antibiotic was added so that a WT positive control could be made. A negative control was also made, with antibiotic. Therefore, in one dish, pluc232A-transformed bacteria were plated and, in a second one, pluc232G-transformed bacteria were plated. The remaining two dishes were plated with WT bacteria. A schematic of this procedure can be found in **Figure 3.4**. All steps were done in sterile conditions. All Petri dishes were incubated over-night at 37°C.



**Figure 3.4 *E. coli* transformation protocol.** This protocol was done under sterile conditions and every LB-agar-containing Petri-dish with bacteria was incubated at 37°C over-night (not shown). WT (JM109 competent non-transformed *E. coli*), pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G).

#### 3.3.1.4 Plasmid amplification, extraction and purification

For bacterial culture growth, colonies were selected (picked) under sterile conditions, from the Petri dishes obtained in subsection 3.3.1.3 (either transformed with pluc232A or pluc232G). Each picked colony was added to a 50-mL centrifuge tube with 5 mL of LB medium and 100 µg/mL of antibiotic (Ampicillin). All tubes were incubated overnight at 37°C with agitation (215 rpm). Positive (without ampicillin) and negative (with ampicillin) controls were also made for a colony from the positive control of the previous subsection.

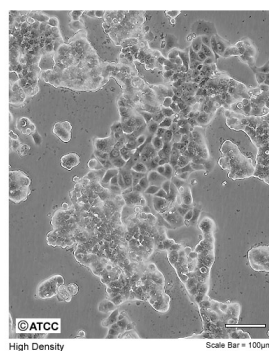
For plasmid extraction and purification, E.Z.N.A.<sup>®</sup> Plasmid DNA kit (Omega Bio-tek, Cat. #D6942-02) was used according to manufacturer's protocol. Briefly, pluc232G and pluc232A plasmids were each obtained in 50 µL of elution buffer. Plasmid DNA concentration was obtained by spectrophotometer nanodrop 2000 and plasmid integrity and size were confirmed by a DNA agarose-TBE 0.6% gel electrophoresis with 1 µg/µL of Green-safe. NZYDNA ladder III was used as a size marker in the electrophoresis. Insert integration was confirmed using the FastDigest *Cai*I (Thermo Scientific<sup>®</sup>, Cat. #FD1394) restriction enzyme according to the manufacturer's protocol, followed by DNA gel electrophoresis (as described above).

#### 3.3.1.5 Cell line

To perform the *in vitro* functional validation, we used the Michigan Cancer Foundation-7 (MCF-7) human breast adenocarcinoma cell line - **Figure 3.5.**

MCF-7 cells were first isolated in 1970 from the breast tissue of a 69-year old Caucasian women (Soule et al., 1973). This cell line retains several characteristics of the differentiated mammary epithelium. Namely, it expresses the oestrogen and progesterone, but not *HER2* (human epidermal growth factor) receptors. Moreover, these cells are suitable for transfection.





**Figure 3.5 MCF-7 cell line.** (Adapted from ATCC)

#### 3.3.1.6 Cell culture

Frozen MCF-7 cells (-150°C) were thawed and cultured in an incubator (Thermo Electron Corporation – Forma Direct Heat CO<sub>2</sub> Incubator) at 37°C with 5% CO<sub>2</sub> in Dulbecco's modified eagle medium (DMEM) supplemented with 10% foetal bovine serum (FBS), 100 units/mL penicillin and 100 µg/mL streptomycin (1% P/S) until they reached 80% confluence in an 80 cm<sup>2</sup> T-flask.

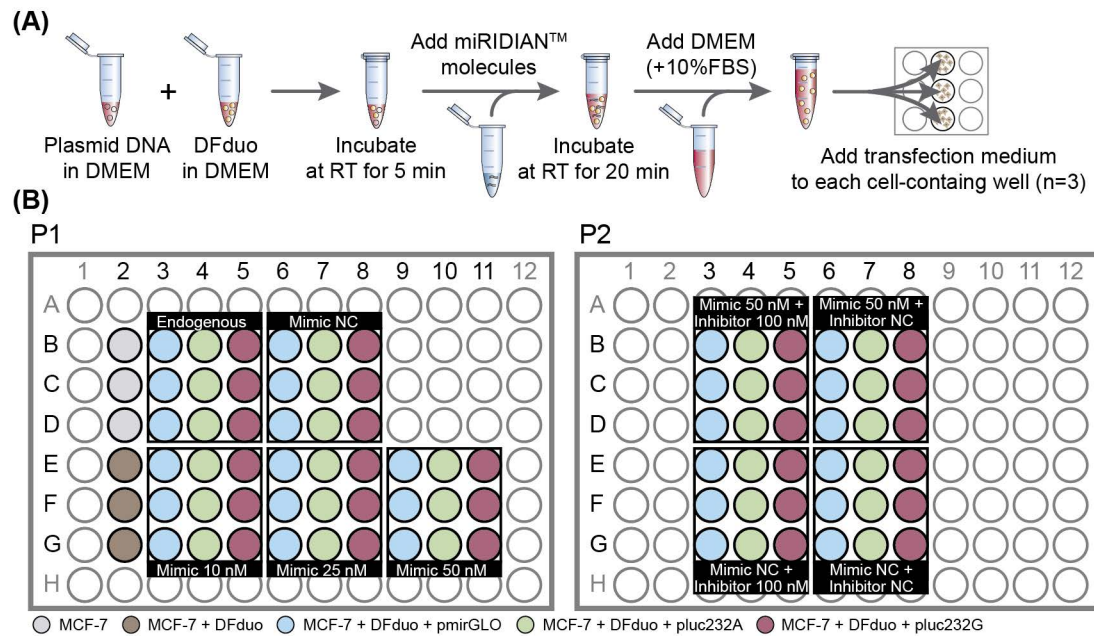
#### 3.3.1.7 Transfection and luciferase assay

A pilot transfection was carried in which  $1.5 \times 10^4$  MCF-7 cells per 96-well were seeded overnight in 75 µL of DMEM (with 10% FBS and 1% P/S). These were transfected with 75 ng of the pmirGLO empty vector or the pmirGLO plasmid containing either one of the two alleles of rs6884232 (pluc232A or pluc232G), using the Polyethylenimine (PEI) transfection reagent (Sigma-Aldrich®) at 0.45 µg/well. Three biological replicates were used. Plasmid DNA was added to DMEM without both FBS and antibiotic, followed by the addition of PEI. The transfection solution was incubated for 10 min at room temperature. Pre-warmed DMEM with 10% FBS and 1% P/S was added and 75 µL of the transfection solution was added to each 96-well. The solution was vortexed briefly between each step. Luciferase activity was measured twenty-four hours or forty-eight hours after transfection using the Dual-Glo Luciferase Assay System (Promega®, Cat. #E2920) according to the manufacturer's protocol with the Infinite M200 microplate reader (Tecan, integration = 500 ms) in a non-sterile white polystyrene 96-well plate (Nunc®).



Similarly to the previous transfection, MCF-7 cells ( $10^4$  per 96-well seeded overnight) were transfected with the pmirGLO empty vector or the pmirGLO plasmid (pluc232A or pluc232G), using the DharmaFECT Duo transfection reagent (Dharmacon™, Cat. #T-2010) at 0.2  $\mu$ L/well according to the manufacturer. Dharmacon's miRIDIAN™ hsa-miR-21-3p mimics (Cat. #C-301023-01-0002), hairpin inhibitors (Cat. #IH-301023-02-0002) and respective negative controls (Cat. #CN-001000-01-05 and #IN-001005-01-05) were co-transfected at diverse concentrations. miRNA mimics are dsRNA molecules that “mimic” the miRNA duplex generated by Dicer. As a result, they are incorporated and processed by Ago, generating a mature synthetic miRNA similar to the endogenous miRNA. miRNA hairpin inhibitors are ssRNA molecules that are the reverse complement of mature miRNAs. Therefore, they are intended to inhibit endogenous miRNAs by irreversibly binding to the mature miRNA-loaded RISC and thus preventing the interaction of the mature miRNA to its endogenous mRNA targets. miRNA mimics and hairpin inhibitors negative controls act by respectively mimicking or inhibiting cel-miR-67. Since cel-miR-64 has a minimal sequence identity with miRNAs in human, it can be used as a negative control.

A transfection negative control (MCF-7 cells) and a mock transfection control (MCF-7 cells with DharmaFECT Duo) were also prepared. Three biological replicates were used. **Figure 3.6** illustrates the co-transfection protocol (panel A) and the 96-well plate design containing the miRIDIAN™ molecules transfection concentrations (panel B). Twenty-four hours after transfection, Dual-Glo Luciferase Assay System (Promega®, Cat. #E2920) was carried according to manufacturer's protocol and luciferase activity was measured with the Infinite M200 microplate reader (Tecan, integration = 500 ms) in a non-sterile white polystyrene 96-well plate (Nunc®). For each well, the ratio of firefly luciferase luminescence to renilla luciferase luminescence was determined.



**Figure 3.6 Luciferase assay transfection protocol.** (A) Co-transfection protocol of the plasmid DNA (pmirGLO or pluc232A or pluc232G) and miRIDIAN™ double-stranded RNA molecules using the DharmaFECT Duo transfection reagent is illustrated. From each transfection medium (last micro-centrifuge tube), three biological replicates were made. (B) Two 96-well plates (P1/2) were used as illustrated. Mimic- and inhibitor- negative controls were used at 100 nM (not shown). Dulbecco's modified eagle medium (DMEM), DharmaFECT Duo transfection reagent (DFduo), room temperature (RT), foetal bovine serum (FBS), negative control (NC), pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G).

### 3.3.1.8 Statistical analysis

Data obtained by Dual-Glo luciferase assay was analysed in R in combination with the *t.test* function from the stats (v3.4.1) R package, containing R statistical functions (R Core Team, 2017). Luciferase ratios were log<sub>2</sub>-transformed to better represent proportional changes in luciferase activity. Under the null hypothesis, that the means of log<sub>2</sub>-transformed luciferase ratios (n=3) between groups are equal, a two-sided Welch's *t*-test was applied to evaluate statistically significant differences within each assayed condition. Welch's *t*-test is a parametric test that does not assume that the variances are equal for normally distributed samples. Standard error of mean (s.e.m.) for log<sub>2</sub>-transformed luciferase ratios was also calculated.

All the statistical graphics were generated using the ggplot2 R package (version 2.2.1), as mentioned earlier.

## **CHAPTER IV**

# **Results**

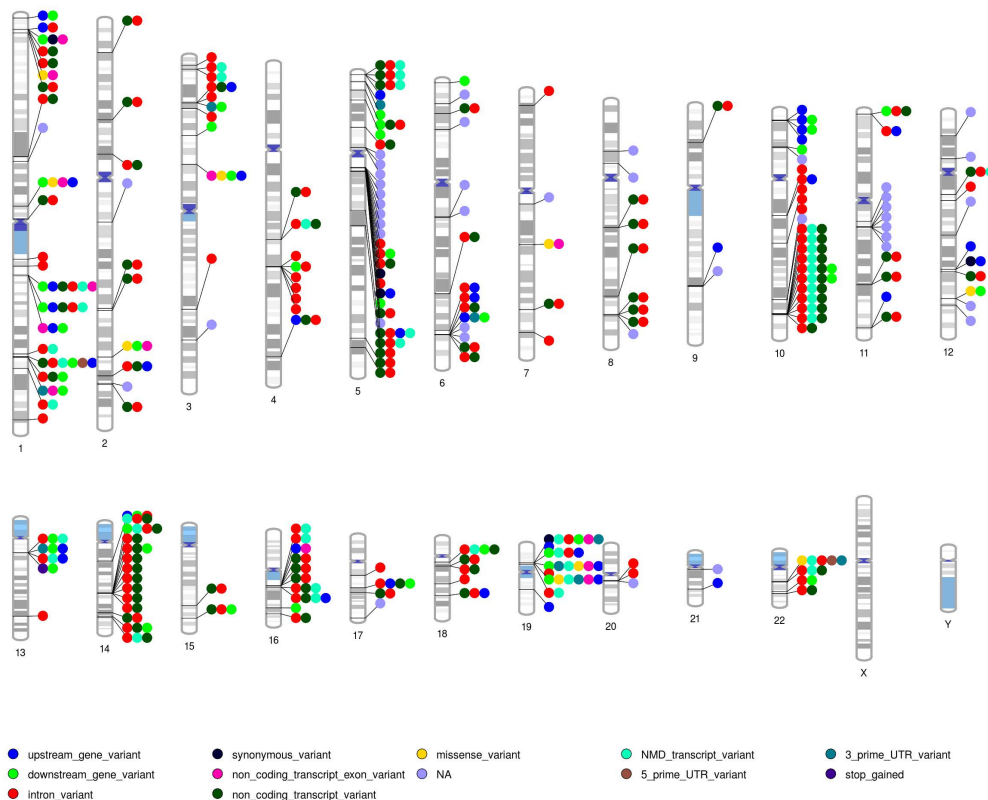


## 4. RESULTS

### 4.1 DATASET OF BREAST CANCER RISK-ASSOCIATED SNPS

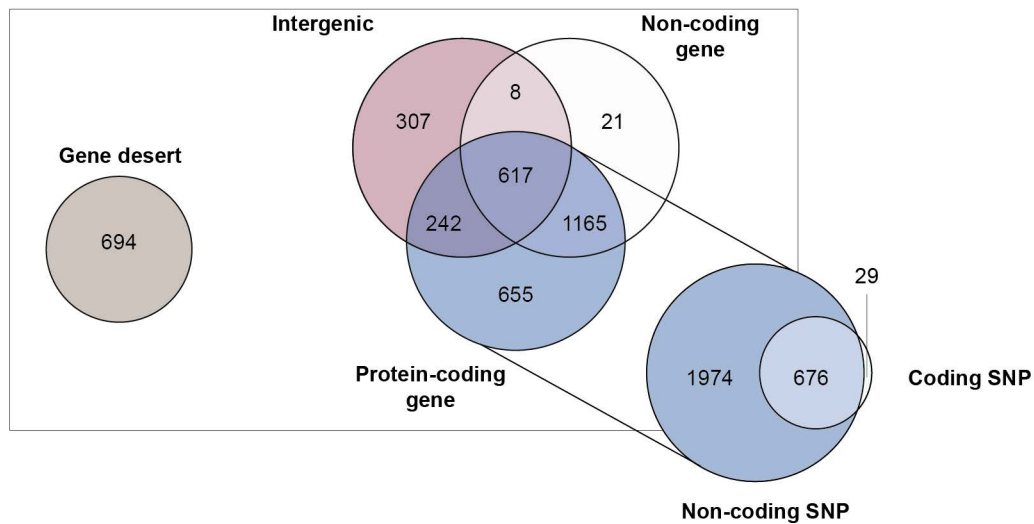
223 GWAS-significant SNPs for BC risk in 108 *loci* were retrieved from public databases and publications (**Annex A**). **Figure 4.1** illustrates their chromosomal localizations. After identifying SNPs in high linkage disequilibrium with those – proxies – (1000 Genomes Project, CEU population,  $r^2 \geq 0.8$ ), our dataset increased to 3891 unique SNPs. Noteworthy, proxies were not found for 21 GWAS-significant SNPs using the above dataset, either because they were not genotyped in the 1000 Genomes Project or there was no information regarding their proxies in that same dataset.

All SNPs were further queried for their annotation in the Ensembl database (release 87), a process which resulted in the automatic loss of 182 SNPs. These SNPs were automatically excluded from the query output of the biomaRt R package because they were flagged by Ensembl for containing errors or inconsistencies in their annotation.



**Figure 4.1 Chromosomal localizations and transcript consequences of BC GWAS-significant SNPs.**

Out of the remaining 3709 SNPs, over one fourth were annotated to “gene deserts” (694) or intergenic regions (307). The remainder were located in either ncRNA genes or PCG’s, for a total of 163 gene symbols (HGNC, HUGO gene nomenclature committee). The vast majority were located in non-coding regions of PCGs, as shown in **Figure 4.2**.



**Figure 4.2** Venn diagram containing the genomic localizations of the 3709 queried SNPs.

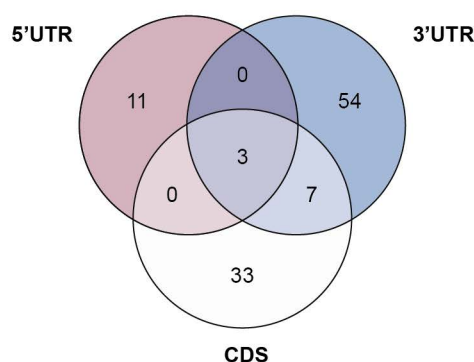
## 4.2 MIRNA BIOGENESIS AND SEED REGION ALTERATION ARE MECHANISMS UNLIKELY INVOLVED IN BREAST CANCER RISK

To start identifying possible mechanisms by which SNPs could be affecting miRNA-mediated regulation in BC risk, we started by assessing how many SNPs in our dataset were located to miRNA genes or nearby. These SNPs would have the putative ability to change the miRNA sequence, thus altering their biogenesis, or target gene list.

Interestingly, none of the SNPs mapped to miRNA genes, suggesting that miRNA biogenesis or alteration of the “seed” sequence are unlikely mechanisms associated with BC risk.

### 4.3 MOST SNPs ASSOCIATED WITH BREAST CANCER RISK ARE LOCATED TO THE NON-CODING REGIONS OF PROTEIN CODING GENES

Next, to identify those risk-associated SNPs that could be altering miRNA-mediated gene regulation via altering of the target sequence to which miRNAs bind, we assessed how many of the GWAS-SNPs located to 5'UTR, CDS and/or 3'UTR of PCGs. We found that 108 SNPs were located within mRNA sequences, and their distribution inside the genes can be found in **Figure 4.3**.



**Figure 4.3** Venn diagram regarding the location of the candidate *cis*-regulatory SNPs within protein-coding genes.

### 4.4 ALLELE-SPECIFIC miRNA TARGET-PREDICTION ANALYSIS

Selected SNPs were then evaluated for putative differential miRNA-binding, induced by the different alleles of each candidate SNP. To achieve this, we started by searching the literature for available miRNA-target prediction algorithms which could perform SNP allele queries. Key differences and similarities between freely-available algorithms are presented in **Table 4.1**.

**Table 4.1 Comparison between the freely-available web-based miRNA-target prediction algorithms.** Selected algorithms are highlighted in grey. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP), untranslated region (UTR), coding sequence (CDS). University of California Santa Cruz (UCSC).

Algorithm	Method	SNP input	Search	Library	Availability	Ref.
<b>TargetScan</b>	Seed complementarity, Conservation and several other features.	No	3'UTR	Gencode-based (hg19)	Online search and source code (Perl language) download	(Agarwal et al., 2015)
<b>miRanda</b>	Complementarity and thermodynamics	No	5'UTR, CDS, 3'UTR	-	Online search and source code (written in C) download	(Enright et al., 2003a)
<b>PicTar</b>	Complementarity and thermodynamics.	No	3'UTR	UCSC genome browser	Online search	(Krek et al., 2005)
<b>DIANA microT-CDS</b>	High-throughput PAR-CLIP sequencing target site overlap and several other features.	No	3'UTR, CDS	Ensembl	Online search	(Paraskevopoulou et al., 2013; Reczko et al., 2012)
<b>RNAHybrid</b>	Thermodynamics	No	3'UTR	-	Online search and source code download	(Krüger and Rehmsmeier, 2006; Rehmsmeier et al., 2004)
<b>PITA</b>	Thermodynamics and site accessibility.	No	3'UTR	UCSC genome browser	Online search and source code (Perl language) download	(Kertesz et al., 2007)
<b>miRTar</b>	Integrated analysis based on TargetScan, miRanda, RNAHybrid and PITA	No	5'UTR, CDS, 3'UTR	Various	Online search	(Hsu et al., 2011)

Based on these characteristics and capabilities, we chose to use two distinct miRNA-target prediction algorithms: TargetScan and miRanda. These algorithms were chosen because they provided an open-source code which could be implemented in R and because they scored each prediction based in several key characteristics involved in miRNA-target binding.



For the 64 SNPs located in the 3'UTR of PCGs, both the TargetScan and miRanda algorithms were applied. The remainder 44 SNPs, located in either the 5'UTR or CDS, were strictly analysed by the miRanda algorithm, as TargetScan did not permit this analysis.

#### 4.4.1 TargetScan analysis

To perform allele-specific miRNA-binding predictions we modified the TargetScan algorithm, in order for the 64 SNPs annotated in 3'UTRs (**Annex B**) to be analysed in an allele-specific way. Since the input 3'UTR sequences were necessarily obtained from TargetScan, nine unique SNPs were excluded because the dataset did not contain 3'UTR sequences for their correspondent gene. Additionally, for those which had available 3'UTR sequences, when we crossed the sequence's chromosomal range coordinates with the SNP coordinate – location control –, ten unique SNPs were excluded from the analysis for not being located within the available sequences. As a result, only 45 unique SNPs were analysed for differential miRNA-binding.

All 45 SNPs caused context++ scores differences for a total of 380 different miRNAs (approximately nine miRNAs per SNP), suggestive that they are causing differential miRNA binding.

Next, we filtered these results to include only miRNAs expressed in normal breast tissue, ending up by excluding four SNPs: rs9321073 (*RNF146*, ring finger protein 146), rs4973768 (*SLC4A7*, solute carrier family 4 member 7), rs3733126 (*ATXN7*, ataxin 7) and rs1132293 (*ASB13*, ankyrin repeat and SOCS box containing 13).

The remainder 41 SNPs were further ranked for previous evidence of their gene being *cis*-regulated in breast tissue. This information was obtained from DAE data generated in our group (Xavier et al., 2016). Three out of 21 different genes in the TargetScan analysis, did not show significant DAE and therefore rs7793861 in *CYP51A1* (cytochrome P450 family 51 subfamily A member 1), rs1778523 and rs2231375 in *CD160* (natural killer cell receptor

BY55), and rs1046025 in *PSMD6* (proteasome 26S subunit, non-ATPase 6) were not further analysed.

At the end of this phase of the analysis, we obtained 37 BC risk-associated SNPs located in the 3'UTRs of 16 PCGs, with prediction for allele-specific miRNA binding. These 37 SNPs corresponded to 19 initial GWAS-significant SNPs, located in 13 independent BC risk *loci* (12% of the initial 108 GWAS *loci*).

#### 4.4.2 miRanda analysis

Parallel to the previous analysis, we also applied the miRanda algorithm to the 108 risk-SNPs annotated either in the 5'UTR, CDS (**Annex C**) or 3'UTR of PCG's to perform allele-specific miRNA-binding predictions.

All 108 risk-SNPs generated minimum free energy differences for a total of 2257 unique miRNAs.

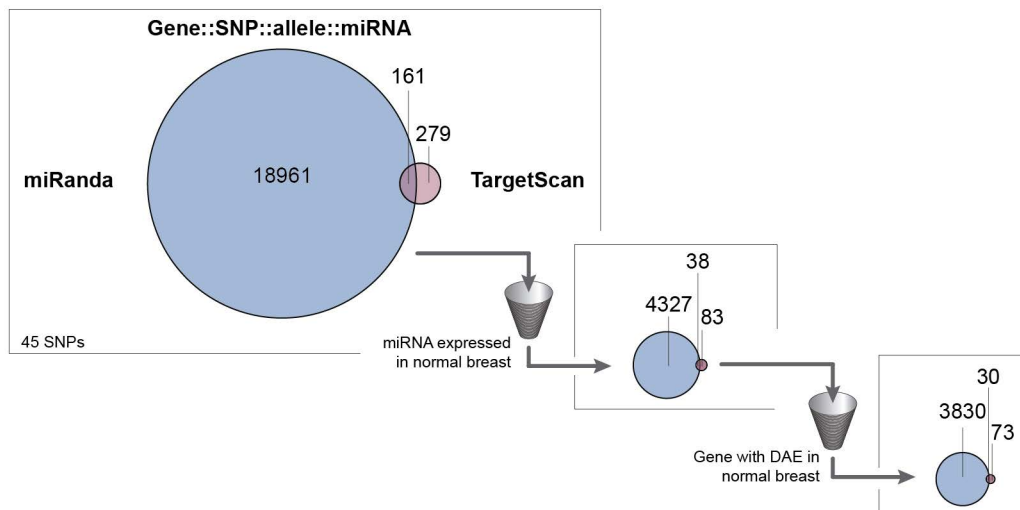
Next, we applied a filter for miRNAs expressed in normal breast tissue as before. Similarly to TargetScan, rs9321073 (located in *RNF146*) did not cause the allele-specific binding of any miRNA expressed in normal breast tissue. Also, rs12654125 (located in *MIER3*, mesoderm induction early response 1 family member 3) did not pass this filter. Both these SNPs were located exclusively in the 3'UTR of their correspondent genes.

The remaining 106 risk-SNPs corresponded to 52 different genes. Similarly to the previous analysis, we filtered those genes for previous evidence of *cis*-regulation through DAE data in normal breast tissue. Three of the genes, were not tested for DAE and, therefore, we cannot draw conclusions regarding whether they are being *cis*-regulated or not. To these genes correspond rs17051298 (located in *TNIP3*, TNFAIP3 Interacting Protein 3), rs743605 (located in *DLX2*, distal-less homeobox 2) and rs2736098 (located in *TERT*, telomerase reverse transcriptase). For the remainder 49 genes, 42 showed significant DAE, suggestive that at least 93 risk-SNPs could be the true causal *cis*-regulatory variant via alteration of miRNA-mediated regulation.

In total, using the two algorithms, we identified 27 independent BC risk *loci* (25% of 108 initially included in the study), located in either 5'UTRs, CDSs or 3'UTRs of PCGs, which showed prediction of differential allelic regulation by miRNAs.

#### 4.4.3 Candidate prioritization

To validate our findings, we prioritized the candidate SNPs for functional characterization, by combining TargetScan' and miRanda' predictions. Out of the 45 common analysed SNPs, miRanda generated almost 44 times more allele-specific miRNA-binding predictions than TargetScan. Moreover, only 30 miRNA-target were common to 16 unique SNPs putatively targeted by miRNAs expressed in breast, and locating in PCGs with evidence of being *cis*-regulated – **Figure 4.4.**



**Figure 4.4** Venn diagram showing the relation between the total number of predictions obtained from the miRanda and TargetScan algorithms for an initial universe of 45 SNPs located in 3'UTRs.

As a consequence, we decided to prioritize SNPs for functional validation primarily based on TargetScan' results because this is a more stringent algorithm, and therefore, there is a higher probability of selecting a true *cis*-regulatory SNP for functional validation.

To help us further establishing this prioritization, we ran the modified version of the TargetScan algorithm for published SNPs with reported functional validation. This would guide us in determining a context++ score threshold. A first example, is that of rs4225 in *APOC3* (apolipoprotein C3), whose T allele specifically binds miR-4271, as reported and functionally validated by Hu and colleagues. The predicted TargetScan context++ score for the T allele of this SNP is of -0.172, with no binding predicted for the C allele.

Another reported example, by Lee and Park, 2016, is that of the specific binding of miR-4273-5p to the A allele of rs7930 in *TOMM20* (translocase of outer mitochondrial membrane 20), for which we obtained a context++ score of -0.151.

Therefore, after selecting a cut-off of -0.151, based on both examples above, we identified six top candidate SNPs with the potential to affect miRNA-binding and produce in vitro significant effects. These variants were rs1573 (*ASB13*), rs2385088 (*ISYNA1*, *inositol-3-phosphate synthase 1*), rs1019806 and rs6884232 (*ATG10*), rs4808616 (*ABHD8*, abhydrolase domain containing 8), and rs3734805 (*CCDC170*, coiled-coil domain containing 170). **Figure 4.5** illustrates the plotting of TargetScan' context++ scores for those variants and for miRNAs expressed in normal breast predicted to bind them.

Next, we decided to focus in the variants located in the 3'UTR of *ATG10* because:

- (i) they made two of our top six results,
- (ii) both variants are in complete LD with the GWAS-significant SNP (rs7707921), which was
- (iii) only recently associated to BC and therefore, the true causal variant is not yet identified, and
- (iv) this *locus* was already being functionally tested in our laboratory, for *cis*-regulatory mechanisms (altered TF binding).

*ATG10* is an E2-like enzyme involved in two ubiquitin-like modifications essential for the autophagosome formation (Nemoto et al., 2003). The

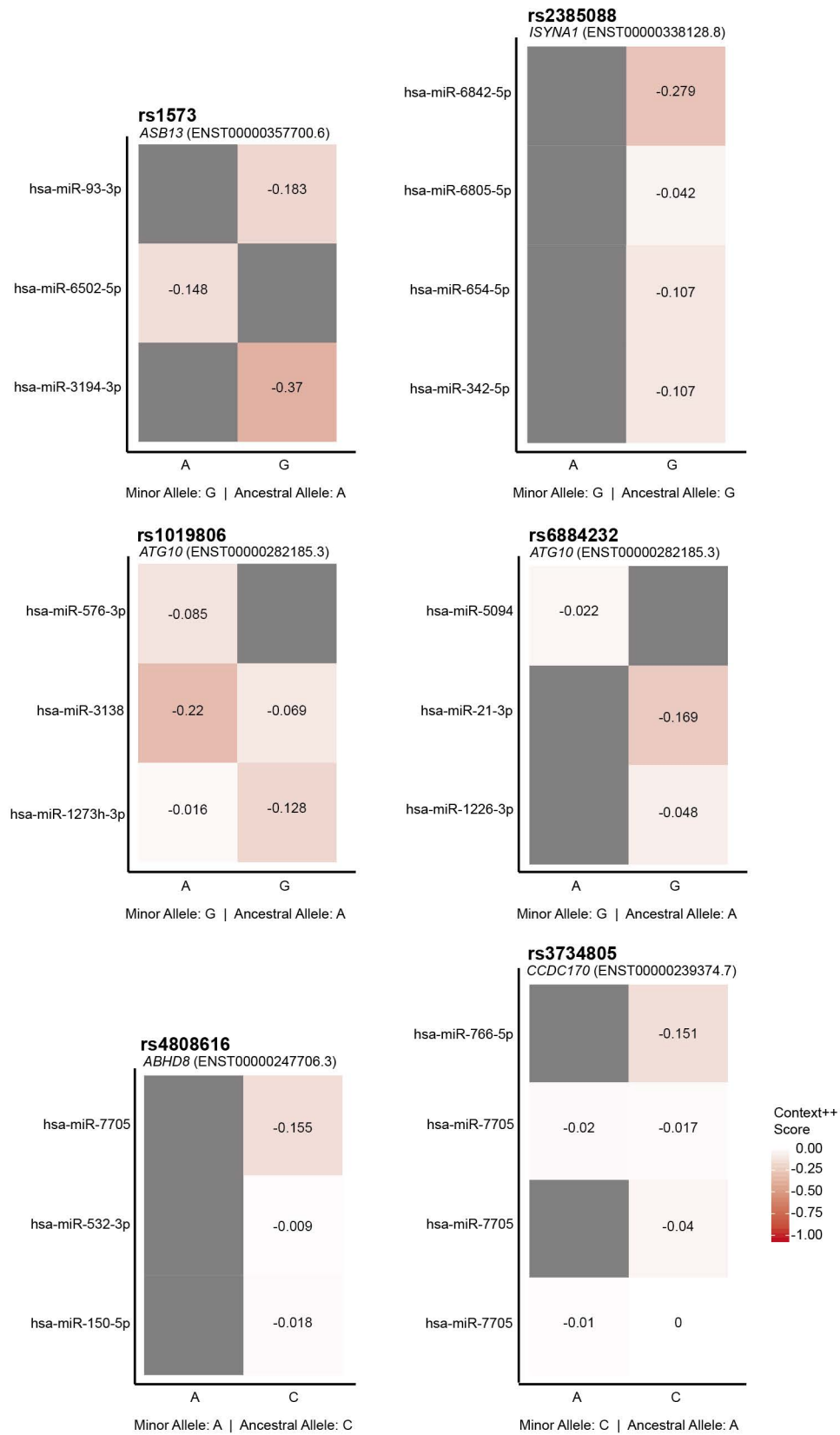
autophagosome is the main structure of macroautophagy, the process by which intracellular material is degraded and recycled.

We decided to carry an in vitro functional validation for the putative allele-specific binding of the passenger strand of hsa-mir-21 to the G allele rs6884232 for three main reasons:

- (i) it was predicted both by TargetScan and miRanda that hsa-miR-21-3p would only bind the G allele, and not the A allele;
- (ii) this SNP is located in the beginning of the 3'UTR, where miRNAs are known to have a more powerful repressing effect; and
- (iii) hsa-miR-21-3p is highly expressed in human breast tumours (based on data from the METABRIC project - EGAD00010000438).

**Table 4.2 Common TargetScan and miRanda predictions for miRNAs expressed in normal breast tissue.** Results are ordered based on 1) increasing context++ score and 2) increasing minimum free energy (MFE) of allele A. The putative SNP::miRNA interaction selected for functional validation is in bold. In dark pink is the allele which is predicted to bind more strongly a miRNA, whereas in light pink is the allele which is predicted to bind less strongly. In grey are alleles which were not predicted to bind the miRNA. Top six predictions are selected/highlighted by a context++ score  $\leq -0.151$ .

Gene	Chr.	SNP	Allele	A.A.	miRNA	TargetScan						miRanda					
						Allele A		Allele B				Allele A		Allele B			
						Site Type	Context++ Score	A	B	Context++ Score	Site Type	Score	MFE (kcal/mol)	A	B	Score	MFE (kcal/mol)
ASB13	10	rs1573	A/G	A	hsa-miR-3194-3p	8mer-1a	-0.367	G	A	NA	NA	151	-19.17	G	A	NA	NA
ISYNA1	19	rs2385088	A/G	G	hsa-miR-6842-5p	7mer-m8	-0.279	G	A	NA	NA	152	-26.8	G	A	NA	NA
ATG10	5	rs1019806	G/A	A	hsa-miR-3138	7mer-1a	-0.22	A	G	-0.069	6mer	121	-20.18	A	G	121	-20.12
<b>ATG10</b>	<b>5</b>	<b>rs6884232</b>	<b>G/A</b>	<b>A</b>	<b>hsa-miR-21-3p</b>	<b>7mer-m8</b>	<b>-0.169</b>	<b>G</b>	<b>A</b>	<b>NA</b>	<b>NA</b>	<b>159</b>	<b>-18.95</b>	<b>G</b>	<b>A</b>	<b>NA</b>	<b>NA</b>
ABHD8	19	rs4808616	C/A	C	hsa-miR-7705	7mer-m8	-0.155	C	A	NA	NA	149	-17.11	C	A	NA	NA
CCDC170	6	rs3734805	A/C	A	hsa-miR-766-5p	7mer-m8	-0.151	C	A	NA	NA	162	-21.72	C	A	NA	NA
RALY	20	rs8123521	A/C	C	hsa-miR-7854-3p	7mer-m8	-0.133	C	A	NA	NA	146	-19.19	C	A	NA	NA
ATG10	5	rs1019806	G/A	A	hsa-miR-1273h-3p	7mer-m8	-0.128	G	A	NA	NA	148	-18.91	G	A	NA	NA
ISYNA1	19	rs2385088	A/G	G	hsa-miR-342-5p	6mer	-0.107	G	A	NA	NA	120	-17.48	G	A	NA	NA
ISYNA1	19	rs2385088	A/G	G	hsa-miR-654-5p	6mer	-0.107	G	A	NA	NA	120	-16.54	G	A	NA	NA
SLC4A7	3	rs1051545	T/C	C	hsa-miR-616-3p	7mer-1a	-0.092	T	T	NA	NA	135	-16.41	T	T	NA	NA
ATG10	5	rs1019806	G/A	A	hsa-miR-576-3p	7mer-m8	-0.084	A	G	NA	NA	150	-16.56	A	G	NA	NA
CCDC170	6	rs3734806	G/A	G	hsa-miR-219b-5p	7mer-m8	-0.077	A	G	NA	NA	155	-20.18	A	G	NA	NA
MARCH6	5	rs1287599	C/A/G	A	hsa-miR-29b-2-5p	7mer-1a	-0.072	A	C/G	NA	NA	122	-16.94	A	C/G	NA	NA
RNF115	1	rs12123298	G/A/C	G	hsa-miR-642a-5p	7mer-1a	-0.068	G	C/G	NA	NA	106	-16.17	G	C/G	NA	NA
ATG10	5	rs6884232	G/A	A	hsa-miR-1226-3p	6mer	-0.048	G	A	NA	NA	120	-16.35	G	A	NA	NA
ISYNA1	19	rs2385088	A/G	G	hsa-miR-6805-5p	6mer	-0.042	G	A	NA	NA	126	-21.77	G	A	NA	NA
RNF115	1	rs12123298	G/A/C	G	hsa-miR-6511a-3p	6mer	-0.028	G	A/G	NA	NA	82	-19.8	G	A/G	NA	NA
MDM4	1	rs10900597	C/T	C	hsa-miR-589-3p	7mer-m8	-0.02	T	C	NA	NA	158	-22.88	T	C	NA	NA
RALY	20	rs6119447	A/G	G	hsa-miR-4286	6mer	-0.017	G	A	NA	NA	120	-16.34	G	A	NA	NA
RALY	20	rs8123521	A/C	C	hsa-miR-4779	6mer	-0.014	C	A	NA	NA	127	-20.13	C	A	NA	NA
SSBP4	19	rs10442	A/C/G/T	C	hsa-miR-664a-3p	7mer-1a	-0.01	G	A/C/T	NA	NA	135	-21.59	G	A/C/T	NA	NA
ADAMTS16	5	rs3806872	C/A/T	C	hsa-miR-3157-5p	6mer	-0.01	A	C/T	NA	NA	128	-21.17	A	C/T	NA	NA
ABHD8	19	rs4808616	C/A	C	hsa-miR-532-3p	6mer	-0.009	C	A	NA	NA	128	-19.41	C	A	NA	NA
ELL	19	rs1043327	A/G	G	hsa-miR-3150b-3p	6mer	0	G	A	NA	NA	143	-22.51	G	A	NA	NA
RNF115	1	rs12123298	G/A/C	G	hsa-miR-5006-3p	6mer	0	G	A/C	NA	NA	122	-21.43	G	A/C	NA	NA
ADAMTS16	5	rs3806872	C/A/T	C	hsa-miR-500b-3p	6mer	0	T	C/A	NA	NA	125	-21.12	T	C/A	NA	NA
CCDC170	6	rs3734805	A/C	A	hsa-miR-3928-3p	6mer	0	C	A	NA	NA	125	-18.61	C	A	NA	NA
RNF115	1	rs12123298	G/A/C	G	hsa-miR-4755-5p	6mer	0	G	A/C	NA	NA	116	-16.12	G	A/C	NA	NA



**Figure 4.5 Top six predictions.** Plots containing TargetScan' predictions are shown for microRNAs expressed in breast.

#### 4.5 FUNCTIONAL VALIDATION OF rs6884232 DIFFERENTIAL ALLELIC BINDING TO hsa-mir-21-3p

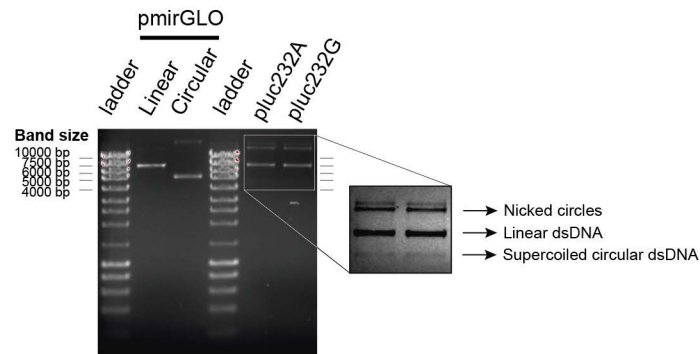
One of the top context++ scores that we identified for a risk-associated SNP, was that of the G allele of rs6884232 (context++ score = -0.169), located in the 3'UTR of ATG10, with predicted allele-specific binding to hsa-miR-21-3p. To functionally validate this result, we used a dual-luciferase system, in which it would be expected the reduction of luciferase activity in the presence of the G allele due to increased miR-21-2p binding and consequential translational repression.

First, the pmirGLO vector was linearized with both *Xho*I and *Xba*I restriction enzymes. Linearization confirmation was performed by agarose gel electrophoresis. As shown in **Figure 4.6**, the second and the third lanes of the gel were loaded with the linearized and circular pmirGLO vector, respectively. The linearized pmirGLO vector is located slightly below the 7500 bp size mark, correctly identifying its size (7342 bp after linearization by *Xho*I and *Xba*I).

After reconstituting the oligonucleotide pairs containing the A or the G allele of rs6884232 in nuclease free-water, and annealing them, we ligated each annealed oligonucleotide pair to the pmirGLO vector. Ligation was verified through agarose gel electrophoresis, shown in **Figure 4.6**. For both ligations, a faint band corresponding to supercoiled circular dsDNA was generated (**Figure 4.6**, panel amplification of the last two lanes), suggesting the correct insert-vector ligation, both at the *Xho*I and *Xba*I overhangs.

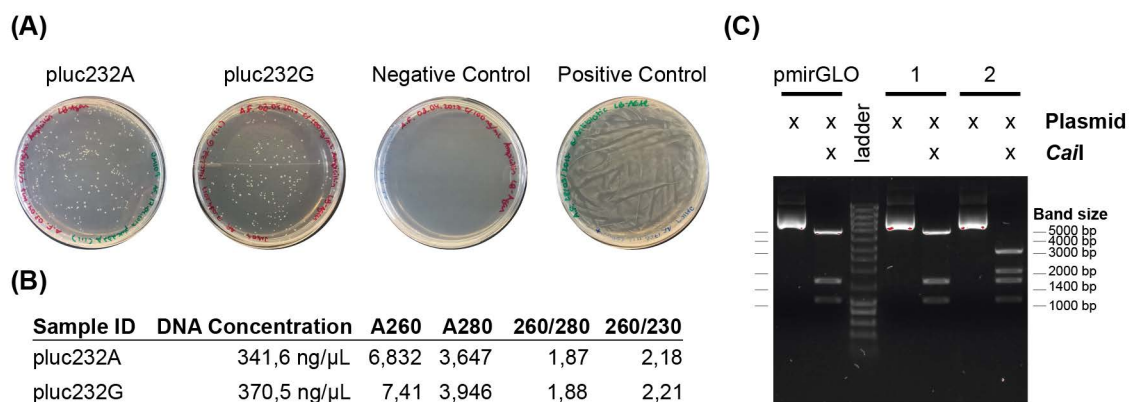
Next, we proceeded to the bacterial transformation of pluc232A and pluc232G. Both transformed bacteria grew in ampicillin-containing medium (**Figure 4.7-A**), due to the expression of  $\beta$ -lactamases encoded in the vector backbone. Correspondingly, WT bacteria without endogenous ampicillin resistance did not grow in ampicillin-containing medium (negative control), whereas substantial growth was observed in antibiotic-free medium (positive control).





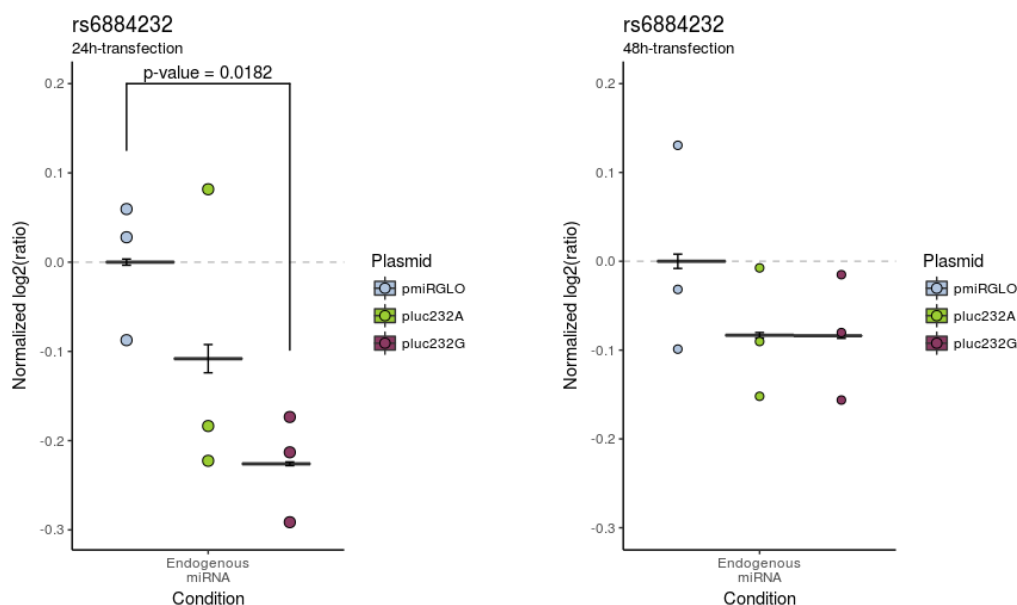
**Figure 4.6 DNA gel electrophoresis for pmirGLO linearization and ligation assessment.** pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G), double-stranded (ds).

Several colonies were picked from each plasmid-transformed dish. After DNA extraction and purification, a sample of each colony DNA was linearized using the *CaII* restriction enzyme to confirm insert integration into the vector backbone. As shown in **Figure 4.7-C**, the pmirGLO vector only contains three restriction sites for *CaII*, generating three linear dsDNA bands after separation in an agarose gel electrophoresis. Plasmid DNA extracted from each transformed bacterial colony can either contain the insert or not. If it contains the insert (either pluc232A or pluc232G), digestion with *CaII* will generate an additional band visible after the electrophoresis in agarose gel (well 1 vs 2 in **Figure 4.7-C**).



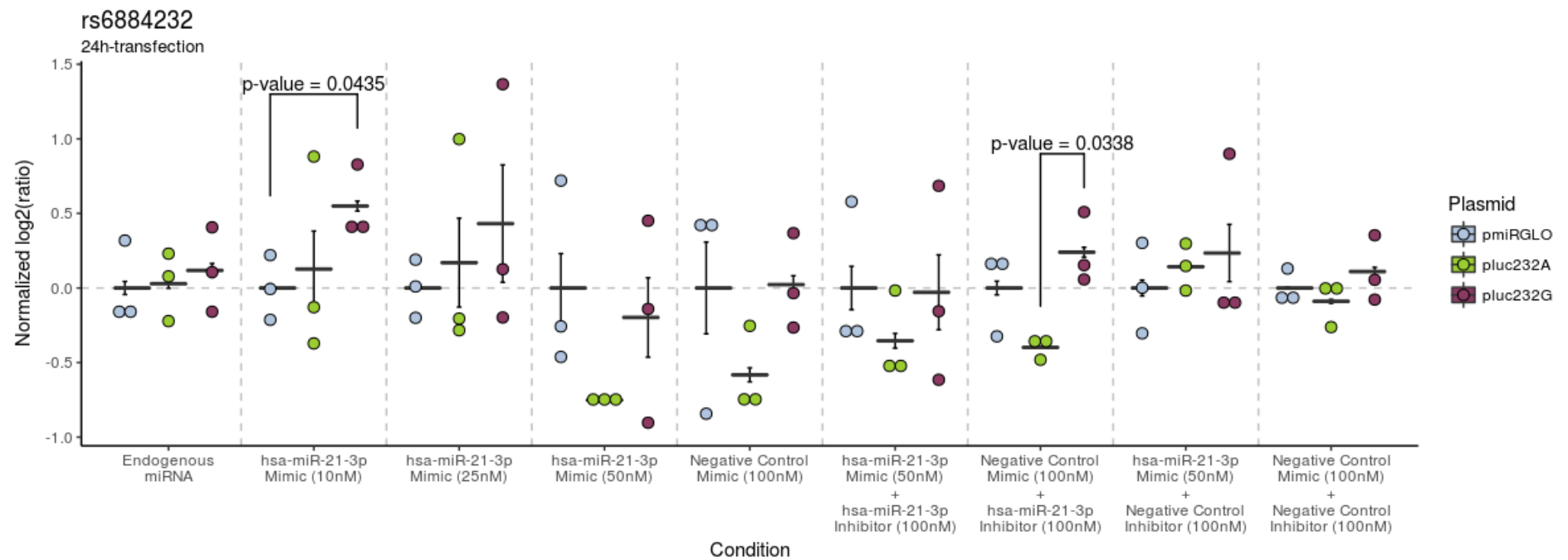
**Figure 4.7 Transformation and insert integration confirmation.** (A) pluc232A- and pluc232G-transformed *E. coli*, and respective negative and positive control. (B) DNA quantification by spectrophotometer. (C) Insert confirmation assessment by DNA agarose-gel electrophoresis after *CaII* enzymatic digestion. Plasmids containing the insert generate four bands (2), whereas plasmids which do not contain the insert only generate three bands (1). pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G),

In order to evaluate the allele-specific binding of hsa-miR-21-3p to the G allele of rs6884232, we transfected human breast adenocarcinoma immortalised cells (MCF-7) with either the pmirGLO vector, pluc232A and pluc232G, and evaluated luciferase activity 24h and/or 48h after transfection. Results can be found in **Figure 4.8**. No significantly statistical differences were observed between the alleles (pluc232A vs pluc232G).



**Figure 4.8 Dual-Glo Luciferase Assay for rs6884232.** Luciferase activity was measured 24h or 48h after transfection. Horizontal bars represent the arithmetic mean of log<sub>2</sub>-transformed ratios, whereas the dots represent the log<sub>2</sub>-transformed ratios (n=3). Results were further subtracted by the arithmetic mean of log<sub>2</sub>-transformed ratios of pmirGLO-transformed cells. Error bars are represented as mean ± standard error of mean (s.e.m.). Pair-wise Welch two-sided t-test was applied. Statistical significance was considered for a p-value < 0.05. pmirGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmirGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G).

Next, we evaluated luciferase activity in combination with miRNA mimics, inhibitors and respective negative controls. Results for pmirGLO-, pluc232A- and pluc232G-transfected cells can be found in **Figure 4.9**. No statistically significant differences were obtained between alleles when hsa-miR-21-3p mimics were added, suggesting that this miRNA does not bind differentially in the presence of the A or G allele of rs6884232. However, our results show high variance amongst biological replicates. Unexpected statistically significant differences between alleles were obtained for the combinatorial use of the mimic negative control and hsa-miR-21-3p inhibitors.



**Figure 4.9 Dual-Glo Luciferase Assay for rs6884232 and hsa-miR-21-3p.** Luciferase activity was measured 24h after transfection. Horizontal bars represent the arithmetic mean of log<sub>2</sub>-transformed ratios, whereas the dots represent the log<sub>2</sub>-transformed ratios (n=3). Results were further subtracted by the arithmetic mean of log<sub>2</sub>-transformed ratios of pmiRGLO-transformed cells. Error bars are represented as mean  $\pm$  standard error of mean (s.e.m.). Pair-wise Welch two-sided t-test was applied. Statistical significance was considered for a p-value < 0.05. pmiRGLO plasmid containing the oligonucleotide pair with the A allele of rs6884232 (pluc232A), pmiRGLO plasmid containing the oligonucleotide pair with the G allele of rs6884232 (pluc232G).



# **CHAPTER V**

## **Discussion**



## 5. DISCUSSION

In this study, we performed *in silico* analysis to assess the contribution of miRNA-mediated *cis*-regulatory SNPs to BC risk. For the first time, we combine BC GWAS and DAE data in normal breast tissue with miRNA regulation, and reveal the existence of putative 27 risk *loci* (out of 108) to be differentially regulated by miRNAs. Also, we carried *in vitro* functional validation for one of the best candidates identified. We evaluated the effect of the putative specific binding of hsa-miR-21-3p to the G allele of rs6884232 (located at the 3'UTR of *ATG10*) on *ATG10* expression. However, no statistically significant differences on *ATG10* expression were observed in the presence of the two alleles of this SNP.

### 5.1 LOCALIZATION OF BC RISK SNPS

One striking GWAS observation is that most risk-associated SNPs identified for common complex traits and diseases are located in non-coding regions (MacArthur et al., 2017), suggesting that they may be conferring risk by regulating gene expression levels. The few functional studies already performed for these risk variants, showed them to be *cis*-regulatory, mainly by altering TF binding. However, it is our belief that other important *cis*-regulatory mechanisms are being overlooked in these studies. To identify possible candidate *cis*-regulatory SNPs that act by altering miRNA post-transcriptional regulation, we retrieved BC risk-SNPs identified through GWAS and their genomic localizations. As mentioned above, most BC risk-SNPs were located in non-coding regions.

Since miRNA target recognition in mammals largely occurs by partial complementarity, we hypothesised that the link between miRNAs and SNPs would primarily involve miRNA biosynthesis rather than its binding to a given target gene. Likewise, 50% of all annotated human miRNA genes are found in fragile or cancer-associated sites (Calin et al., 2002, 2004). However, in our analysis of BC susceptibility *loci*, we did not find SNPs located within miRNA

genes. This suggests that miRNA biogenesis and target-binding alteration are mechanisms not likely to be involved in BC risk.

The lack of BC risk SNPs within miRNA genes could be due to our “biased” analysis for SNPs in high LD with GWAS-significant SNPs in a population of European ancestry. In fact, there are several examples of miRNA SNPs associated with BC, including in other populations (Chacon-Cortes et al., 2015; Hoffman et al., 2009; Kabirizadeh et al., 2016; Kontorovich et al., 2010; Smith et al., 2012).

Also, there are still several unidentified BC risk *loci*, accounting for approximately half of the genetic component of the disease (missing heritability). So, future studies identifying further risk *loci*, may include some that will locate to miRNA genes, and thus our type of analysis should be repeated for any new risk-*loci*.

Of note, 19 SNPs were annotated as downstream or upstream variants of miRNA genes and therefore, they should have been candidates for exerting a *cis*-regulatory function at transcriptional level. However, none of the miRNA genes showed DAE, suggestive that they are not being *cis*-regulated (data not shown).

From the SNPs located in possible miRNA targets, we found at least 108 in 52 genes located across 30 *loci* (reported through 52 GWAS-significant SNPs). This suggests that the associated risk of ~28% of GWAS-identified *loci* could be conferred via differential binding of miRNAs. Interestingly, most of those SNPs are located in the 3'UTR of PCG's, where miRNAs are known to bind preferentially.

## 5.2 COMPARISON OF miRNA-TARGET PREDICTION ALGORITHMS

Given the overall importance and interest in clarifying the biological mechanisms affected by BC risk variants, we would expect the existence of available tools that could perform, in this case, allele-specific miRNA-binding



predictions. However, and to our surprise, none of the available algorithms directly allowed this kind of analysis. Therefore, we selected the TargetScan and miRanda algorithms and developed a systematic pipeline using the R programming language to evaluate the effect of mRNA-located SNPs in miRNA binding.

TargetScan is a more stringent algorithm based on canonical binding of miRNAs to 3'UTRs, whereas miRanda is more flexible algorithm which allows both canonical and non-canonical miRNA binding to the mRNA sequence. Although we ultimately relied more on TargetScan results, both analysis were complementary to each other.

First, we were able to assess allele-specific miRNA binding in the 5'UTR, CDS and 3'UTR of PCG's. Although the miRanda algorithm could perform this directly, TargetScan provides a more reliable prediction by being more stringent in its predictions and thus generating fewer hits for 3'UTR-located variants.

Conversely, being too strict in miRNA-target prediction, could result in the loss of functional miRNA targets which do not have canonical miRNA binding. For instance, Nicoloso and colleagues functionally validated two coding SNPs, one in *TGFB1* (transforming growth factor beta 1) and another in *XRCC1* (X-ray repair cross complementing 1), to cause differential non-canonical miRNA binding (Nicoloso et al., 2010) supporting the importance of analysing also non-canonical miRNA binding.

### 5.3 EXPRESSION AND *CIS*-REGULATION IN BREAST

To identify *cis*-regulatory variants that are conferring BC risk, we began by selecting a list of BC-associated SNPs reported through GWAS. Although these studies provide the statistical association of these variants to BC, they do not point the causal variant neither guarantee a *cis*-functional role.

Several miRNAs exhibit tissue-specific expression (Lim et al., 2005; Mansfield et al., 2004; Wheeler et al., 2006; Zhu et al., 2014). As such, we found it important to filter our results for miRNAs expressed in breast. This

could have been done by using the information from our microarray breast tissue data used to perform DAE analysis. If the expression of the miRNA transcribed SNP was detected, then this miRNA would be expressed in normal breast tissue. However, the microarray used in that experience was mostly coding gene-centred and did not cover the great majority of miRNA genes. Thus, we had to resort to databases which contained miRNA expression data such as miRmine (Panwar et al., 2017).

We filtered our allele-specific miRNA-binding predictions by using miRNA-sequencing data available in the public database miRmine. For clarity, we defined that miRNAs with expression values different than zero would be classified as expressed. This requires careful attention, as some miRNAs might have very low expression levels and therefore not exert observable effects resulting in an overestimation of putative mRNA-miRNA binding sites by our analysis. We also obtained miRNA expression data in BC from the METABRIC miRNA landscape study (after permission). However, this study was based on an older version of miRBase (v16.0) and there were several miRNAs posteriorly identified which were missing in the dataset. As a result, we used this data as complementary information when selecting a SNP::miRNA combination for functional validation.

DAE is a hallmark of *cis*-regulation. Filtering the genes of our miRNA-binding predictions for evidence of DAE in normal breast tissue provides a direct support that the gene is being *cis*-regulated in the breast. Even if the regulatory mechanism is not mediated by miRNAs, or is a combination of several mechanisms, the combination of BC GWAS and DAE in normal breast is likely to provide/identify the true causal variant. Moreover, some genes also exhibit tissue-specific expression (Zhu et al., 2016). Therefore, this filter also contributes to strengthen the evidence that the gene is being expressed in breast tissue.

From these studies, we identified at least 27 BC risk *loci*, that may be conferring risk by altering miRNA post-transcriptional regulation. Moreover, this confirms the importance of studying the effect of genetic variation in other

regulatory mechanisms, besides that of TF binding, as a means to identify risk variants responsible for complex diseases.

Noteworthy, our TargetScan-based *in silico* analysis is corroborated by the already existing functional validation of rs4245739 in *MDM4* (Human homolog of double minute 4, P53-Binding Protein; Wynendaele et al., 2010) and rs11540855 in *ABHD8* (Li et al., 2017) to cause allele-specific miRNA binding.

Our predictions identified the specific-binding of hsa-miR-191-5p to the C allele of rs4245739 with a context++ score of -0.309. Wynendaele and colleagues showed that luciferase activity was significantly reduced when the cells were transfected with reporter constructs containing the C allele, in a cell line which expressed high levels of hsa-miR-191-5p.

For rs11540855 in *ABHD8*, we obtained a predicted context++ score of -0.225 for the specific-binding of hsa-miR-4707-3p to the G allele of this SNP. In the presence of hsa-miR-4707-3p mimics, luciferase activity significantly reduced for the G allele, but not the A allele of rs11540855 (Li et al., 2017).

This supports that our predictions might also be functional.

## 5.4 SNP CANDIDATE PRIORITIZATION FOR FUNCTIONAL VALIDATION

Based on the results from both TargetScan and miRanda, we hypothesized that combining both prediction algorithms would result in a higher probability of identifying a functional miRNA-mediated *cis*-regulatory SNP.

The combined results identified at least six BC risk-associated SNPs candidates to be miRNA-mediated *cis*-regulatory variants, and therefore possible responsible for the observed DAE in these genes, or at least contributing towards it in a more complex pattern of *cis*-regulation (i.e. several contributing mechanisms). These variants were rs1573 (located in *ASB13*), rs2385088 (located in *ISYNA1*), rs1019806 and rs6884232 (located in *ATG10*), rs4808616 (located in *ABHD8*), and rs3734805 (located in *CCDC170*).

*ASB13* (10p15.1) is involved in protein ubiquitination and subsequent proteasomal degradation. Our prediction is that the G allele of rs1573 could be a target for hsa-miR-3194-3p, which would result in the decreased expression of *ASB13* in GG homozygotes and higher expression (in comparison) in AA homozygotes, whereas GA heterozygotes would have intermediate levels of *ASB13* expression. As such, decreased expression of *ASB13* could result in impaired proteasomal degradation. Since the ubiquitin/proteasome system is involved in several cell cycle, cell death and DNA repair pathways (Reviewed in Schmidt and Finley, 2014), its impairment could, in turn, promote tumour development. Moreover, rs1573 is reported through the GWAS-significant SNP rs4414128 (Song et al., 2013), where the minor allele C of rs4414128 is associated with risk of BC. Since rs1573 is in high LD ( $r^2=0.832$ ,  $D'=1$ ) with rs4414128, the minor allele G of rs1573 is also associated with BC risk, consistent with our hypothesis.

*ISYNA1* (19p13.11) codes for an inositol-3-phosphate synthase enzyme, involved in the myo-inositol biosynthesis pathway. Myo-inositol was reported to suppress tumour growth *in vivo* (Kassie et al., 2008; Wattenberg and Estensen, 1996) and ectopic *ISYNA1* expression was shown to increase myo-inositol levels, subsequently suppressing tumour cell growth (Koguchi et al., 2016). Moreover, *ISYNA1* is directly induced by the p53 tumour suppressor (Koguchi et al., 2016), further establishing its tumour suppressor role. Putative targeting of the G minor allele of rs2385088 by hsa-miR-6842-5p from our predictions, would result in *ISYNA1* post-transcriptional repression. Consequently, GG homozygotes would have higher risk of tumour development. However, rs2385088 is reported through rs4808801 (Michailidou et al., 2013), where the minor allele G is associated with lower odds of developing BC. Therefore, our predictions would not explain the association in the GWAS according to the current knowledge of its gene/protein function.

ABHD8 (encoded by a gene located at 19p13.11) is an uncharacterized protein. As a result, we cannot predict its role in tumour development. Still, rs4808616 is in complete LD with the BC-associated SNP rs4808075 (Fehrer et al., 2016), for which the minor allele C was associated with increased risk.

Therefore, the putative targeting the common allele C of rs4808616 and subsequent downregulation of ABHD8 could, in theory, confer BC protection.

*CCDC170* (gene located at 6q25.1) was only recently discovered to play an essential role in Golgi-associated microtubule organization and stabilization (Jiang et al., 2017). Dysregulation of *CCDC170* has been suggested to act either as tumour suppressor or oncogenic (Jiang et al., 2017; Veeraraghavan et al., 2014). rs3734805 is reported itself as a GWAS-significant SNP (Fletcher et al., 2011) and by rs12662670 ( $r^2=0.892$ ,  $D'=1$ ; Hein et al., 2012). In both, the minor allele is associated with increased risk of BC. Thus, our predictions for the binding of has-miR-766-5p to the C allele of rs3734805, and subsequent downregulation of *CCDC170*, would be consistent with loss of microtubule organization and increased tumour predisposition.

*ATG10* is a cytoplasmatic protein involved in autophagy (Nemoto et al., 2003). Autophagy is the catabolic process by which cellular material is degraded by lysosomes or vacuoles and subsequently recycled. Its role in cancer is rather complex, having several studies showing both a tumour suppressor and oncogenic role (Brech et al., 2009; Degenhardt et al., 2006; Mathew et al., 2007; Morselli et al., 2009; Sudarsanam and Johnson, 2010). At least, increased expression of *ATG10* has been associated with colorectal cancer (Jo et al., 2012).

## 5.5 rs6884232: A *CIS*-REGULATORY CANDIDATE

Two candidate miRNA-mediated *cis*-regulatory SNPs were located in the 3'UTR of *ATG10*: rs1019806 (A/G) and rs6884232 (A/G). Both SNPs are in complete LD with the GWAS meta-analysis associated SNP rs7707921 (A/T; Michailidou et al., 2015). The minor allele T of rs7707921 is associated with lower odds (odds ratio = 0.94, 95% confidence interval = 0.90-0.98,  $p$ -value=0.00302) of developing BC. As such, the G minor allele of both rs1019806 and rs6884232 are also associated with protection to BC.

Based on the hypothesis that miRNAs induce post-transcriptional silencing and that increased *ATG10* expression is associated with cancer, only the prediction for rs6884232 is consistent with the protective effect associated

with the minor allele of this SNP. The putative specific binding of hsa-miR-21-3p to the G allele of rs6884232 should cause *ATG10* downregulation in GG homozygotes and GA heterozygotes and thus lead to decreased tumour susceptibility.

mir-21 is a well-known oncogene and miR-21-5p is overexpressed in many human cancers (Hatley et al., 2010; Selcuklu et al., 2009; Si et al., 2007), including that of the breast (Iorio et al., 2005; Sempere et al., 2007; Yan et al., 2008). Additionally, only the passenger strand of mir-21 (miR-21-3p) was shown to be a positive regulator of L1CAM (L1 cell adhesion molecule), a cell surface protein associated with poor prognosis in several tumour malignancies (Altevogt et al., 2016; Chen et al., 2013; Van Gool et al., 2016; Nakaoka et al., 2017; Tangen et al., 2017), including breast cancer (Doberstein et al., 2014a). Overexpression of miR-21-3p strongly increased L1CAM expression in several cancer cell lines (Doberstein et al., 2014b), further suggesting a greater potential for the differential binding of this miRNA to rs6884232.

Another corroborating result for the greater regulatory potential of this SNPs is that rs6884232 is also associated with the DAE levels observed for *ATG10* (Xavier et al., 2016). Furthermore, this SNP was also associated to *ATG10* expression levels in normal breast by eQTL studies (Aguet et al., 2016; **Annex E**).

### 5.5.1 Functional validation of RS6884232

We hypothesised that rs6884232 could explain the observable DAE in *ATG10* by causing allele-specific binding of hsa-miR-21-3p and therefore conferring BC risk.

First, we carried a functional study in a breast adenocarcinoma cell line using a dual-luciferase assay system, with constructs containing a portion of the 3'UTR sequence of *ATG10* with either the A or the G allele of rs6884232. In this experiment, we were predicting a higher decrease of luciferase activity in cells transfected with the G allele *versus* the A allele. This would be a strong indicator that a given miRNA (predicted to be hsa-miR-21-3p) was differentially binding the alleles of rs6884232.

However, no statically significant differences were observed in luciferase activity between alleles nor between the empty vector and either one of the alleles, after both 24 and 48h transfections.

There are two main possible explanations for this:

- (i) either miRNAs do not bind either one of the putative miRNA-target sequences or;
- (ii) miRNAs that do bind the putative miRNA-target sequences are not expressed, or have low expression levels, in the used cell line, MCF-7.

Since one of our European collaborators has measured miRNA expression in MCF-7, namely hsa-miR-21-3p, we hypothesized that the endogenous expression of the latter miRNA was not sufficient to exert observable effects. Gain- and loss-of-function experiments using miRNA mimics and inhibitors, respectively, can be used to specifically evaluate the effect of the miRNA of interest even if the miRNA is not endogenously expressed. Thus, we decided to use the latter approach in combination with the luciferase assay to determine the role of hsa-miR-21-3p in the presence of rs6884232. For this experiment, we evaluated luciferase activity only 24h after transfection, as no statistically significant differences were observed in both pilot transfections mentioned above.

Still, no allele-specific binding differences in luciferase activity were observed in the presence of has-miR-21-3p mimics at different concentrations, suggestive that this miRNA was not differentially binding the putative miRNA-target sequences.

We did find the G allele to have significantly higher luciferase activity in the presence of the hsa-miR-21-3p mimics at 10 nanoMolar (nM) when compared to the empty vector (control). This is suggestive that the mimic was stabilizing the G allele, rather than silencing it. Increased levels of the G allele are consistent with what is observed in our DAE data, as well as by eQTL studies, but to our knowledge there are no reports of binding of miRNAs to the 3'UTR causing stabilization of gene expression. Still, no differences were

observed in the presence of the mimic at 25 and 50 nM. These observations, suggest that the mimic could be exerting some regulatory role in the presence of rs6884232; however, its effect is only visible at low concentrations. This alludes to a possible saturation effect in the assayed conditions. Noteworthy, there has been some controversy in the use of miRNA mimics. High concentrations of miRNA mimics may cause unspecific binding, whereas low concentrations may not be sufficient to efficiently suppress target gene expression (Jin et al., 2015).

Additionally, we examined the effect of hsa-miR-21-3p hairpin inhibitors in combination with hsa-miR-21-3p mimics or mimic negative control. From this experiment, we would expect that the luciferase activity would be similar for both alleles and the empty vector. However, we obtained statistical evidence that the mean between the alleles are not equal in the presence of both the hsa-miR-21-3p hairpin inhibitor and the mimic negative control. Thus, either hsa-miR-21-3p inhibition is allowing other miRNAs to bind allele-specifically, or the mimic negative control is exerting some regulatory effect in rs6884232 luciferase activity.

Interestingly, a co-worker from our group, obtained unexpected statically significant differences between alleles of a different SNP when she used the same mimic negative control. This supports our hypothesis that the mimic negative control is responsible for the allelic differences in that condition. This also compromises the conclusions drawn from the experiments in which the mimics were used.

Furthermore, there was also high variability amongst biological replicates in our experiments, and we will have to repeat the assay to increase confidence in our results and draw our conclusions.



# **CHAPTER VI**

## **Conclusions and Future Directions**



## 6. CONCLUSIONS AND FUTURE DIRECTIONS

To date, the underlying mechanisms by which *cis*-acting genetic variation impacts gene expression differences and disease susceptibility remains poorly understood. Here we evaluated the potential role of miRNA-mediated *cis*-regulation in BC risk assessment and found them predicted to affect a quarter of the BC risk-associated loci identified through GWAS. Moreover, we found that BC risk-associated SNPs are not located in miRNA genes and that miRNA-mediated *cis*-regulation is likely be involved in breast cancer risk by altering miRNA target sequences rather than miRNA sequences.

We performed *in vitro* functional characterization of rs6884232, a 3'UTR variant located in *ATG10*, and found no differences in luciferase activity between both alleles of this SNP for MCF-7 endogenously expressed miRNAs. We also assessed the specific role of hsa-miR-21-3p in the presence of both alleles of rs6884232, but also found no differences in luciferase activity. Still, the high variability in biological replicates and unexpected side-effects in control conditions prevent us from drawing definitive conclusions. In the near future, we will repeat these experiments and attempt to further characterize the role of rs6884232 in BC risk.

To our knowledge, we developed the first systematic study integrating GWAS, DAE and miRNA regulation for BC risk assessment. We will further develop this approach by determining the clinical importance of miRNA-mediated *cis*-acting genetic variants in BC development and progression. Our work will help to improve the understanding of BC aetiology and identify new therapy targets.

## 7. REFERENCES

- Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, 1–38.
- Aguet, F., Brown, A.A., Castel, S., Davis, J.R., Mohammadi, P., Segre, A. V, Zappala, Z., Abell, N.S., Fresard, L., Gamazon, E.R., et al. (2016). Local genetic effects on gene expression across 44 human tissues.
- Ahlbom, A., Lichtenstein, P., Malmström, H., Feychting, M., Hemminki, K., and Pedersen, N.L. (1997). Cancer in twins: genetic and nongenetic familial risk factors. *J. Natl. Cancer Inst.* 89, 287–293.
- Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212.
- Altevogt, P., Doberstein, K., and Fogel, M. (2016). L1CAM in human cancer. *Int. J. Cancer* 138, 1565–1576.
- Ameres, S.L., Martinez, J., and Schroeder, R. (2007). Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell* 130, 101–112.
- Ardekani, A.M., and Naeini, M.M. (2010). The Role of MicroRNAs in Human Diseases. *Avicenna J. Med. Biotechnol.* 2, 161–179.
- Arvey, A., Larsson, E., Sander, C., Leslie, C.S., and Marks, D.S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.* 6, 363.
- Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. *Nature* 455, 64–71.
- Balding, D.J., and Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7, 781–791.
- Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell* 136, 215–233.
- Bell, D.W. (1999). Heterozygous Germ Line hCHK2 Mutations in Li-Fraumeni Syndrome. *Science* (80-. ). 286, 2528–2531.
- Betel, D., Koppal, A., Agius, P., Sander, C., Leslie, C., Obad, S., Lindholm, M., Hedtjärn, M., Hansen, H., Berger, U., et al. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* 2010 118 452, 896–899.
- Bohnsack, M.T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* 10, 185–191.
- Brech, A., Ahlquist, T., Lothe, R.A., and Stenmark, H. (2009). Autophagy in tumour suppression and promotion. *Mol. Oncol.* 3, 366–375.
- Brennecke, J., Stark, A., Russell, R.B., Cohen, S.M., and Marks, D. (2005). Principles of MicroRNA–Target Recognition. *PLoS Biol.* 3, e85.
- Brewster, B.L., Rossiello, F., French, J.D., Edwards, S.L., Wong, M., Wronski,

- A., Whiley, P., Waddell, N., Chen, X., Bove, B., et al. (2012). Identification of fifteen novel germline variants in the *BRCA1* 3'UTR reveals a variant in a breast cancer case that introduces a functional *miR-103* target site. *Hum. Mutat.* 33, 1665–1675.
- Bryois, J., Buil, A., Evans, D.M., Kemp, J.P., Montgomery, S.B., Conrad, D.F., Ho, K.M., Ring, S., Hurles, M., Deloukas, P., et al. (2014). Cis and Trans Effects of Human Genomic Variants on Gene Expression. *PLoS Genet.* 10.
- Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10, 1957–1966.
- Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., et al. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15524–15529.
- Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., et al. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U. S. A.* 101, 2999–3004.
- Cantor, S.B., Bell, D.W., Ganesan, S., Kass, E.M., Drapkin, R., Grossman, S., Wahrer, D.C., Sgroi, D.C., and Lane, W.S. (2001). BACH1, a Novel Helicase-like Protein, Interacts Directly with BRCA1 and Contributes to Its DNA Repair Function. *Cell* 105, 149–160.
- Chacon-Cortes, D., Smith, R.A., Haupt, L.M., Lea, R.A., Youl, P.H., and Griffiths, L.R. (2015). Genetic association analysis of miRNA SNPs implicates MIR145 in breast cancer susceptibility. *BMC Med. Genet.* 16, 107.
- Chen, J., and Lindblom, A. (2001). Germline mutation screening of the STK11/LKB1 gene in familial breast cancer with LOH on 19p. *Clin. Genet.* 57, 394–397.
- Chen, D., Zeng, Z., Yang, J., Ren, C., Wang, D., Wu, W., and Xu, R. (2013). L1cam promotes tumor progression and metastasis and is an independent unfavorable prognostic factor in gastric cancer. *J. Hematol. Oncol.* 6, 43.
- Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436, 740–744.
- Cheung, V. (2016). Individual Variation in Gene Expression (Part 1): Vivian Cheung.
- Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.-Y., Morley, M., and Spielman, R.S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33, 422–425.
- Claus, E.B., Risch, N., and Thompson, W.D. (1991). Genetic analysis of breast

cancer in the cancer and steroid hormone study. *Am. J. Hum. Genet.* **48**, 232–242.

Collaborative Group on Hormonal Factors in Breast Cancer (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* (London, England) **358**, 1389–1399.

Connor, F., Bertwistle, D., Mee, P.J., Ross, G.M., Swift, S., Grigorieva, E., Tybulewicz, V.L., and Ashworth, A. (1997). Tumorigenesis and a DNA repair defect in mice with a truncating *Brca2* mutation. *Nat. Genet.* **17**, 423–430.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194.

Davis, E., Caiment, F., and Tordoir, X. (2005). RNAi-Mediated Allelic trans-Interaction at the Imprinted *Rtl1/Peg11* Locus. *Curr. Biol.* **15**, 743–749.

Degenhardt, K., Mathew, R., Beaudoin, B., Bray, K., Anderson, D., Chen, G., Mukherjee, C., Shi, Y., G  linas, C., Fan, Y., et al. (2006). Autophagy promotes tumor cell survival and restricts necrosis, inflammation, and tumorigenesis. *Cancer Cell* **10**, 51–64.

Doberstein, K., Milde-Langosch, K., Bretz, N.P., Schirmer, U., Harari, A., Witzel, I., Ben-Arie, A., Hubalek, M., M  ller-Holzner, E., Reinold, S., et al. (2014a). L1CAM is expressed in triple-negative breast cancers and is inversely correlated with Androgen receptor. *BMC Cancer* **14**, 958.

Doberstein, K., Bretz, N.P., Schirmer, U., Fiegl, H., Blaheta, R., Breunig, C., M  ller-Holzner, E., Reimer, D., Zeimet, A.G., and Altevogt, P. (2014b). MiR-21-3p is a positive regulator of L1CAM in several human carcinomas. *Cancer Lett.* **354**, 455–466.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. *Genes* (Basel). **504**, 504–511.

Duan, R., Pak, C., and Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum. Mol. Genet.* **16**, 1124–1131.

Dunning, A.M., Healey, C.S., Baynes, C., Maia, A.T., Scollen, S., Vega, A., Rodr  guez, R., Barbosa-Morais, N.L., Ponder, B.A.J., Low, Y.L., et al. (2009). Association of *ESR1* gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.* **18**, 1131–1139.

Dunning, A.M., Michailidou, K., Kuchenbaecker, K.B., Thompson, D., French, J.D., Beesley, J., Healey, C.S., Kar, S., Pooley, K.A., Lopez-Knowles, E., et al. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate *ESR1*, *RMND1* and *CCDC170*. *Nat. Genet.* **48**, 374–386.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and

Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191.

Easton, D.F. (1994). Cancer risks in A-T heterozygotes. *Int. J. Radiat. Biol.* 66, S177-82.

Easton, D.F. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Res.* 1, 14–17.

Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D.P., Thompson, D., Ballinger, D.G., Struwing, J.P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447, 1087–1093.

Ebert, M.S., and Sharp, P.A. (2012). Roles for MicroRNAs in conferring robustness to biological processes. *Cell* 149, 505–524.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003a). MicroRNA targets in *Drosophila*. *Genome Biol.* 5, R1.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003b). MicroRNA targets in *Drosophila*. *Genome Biol.* 5, R1.

Fachal, L., and Dunning, A.M. (2015). From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr. Opin. Genet. Dev.* 30, 32–41.

Fareh, M., Yeom, K.-H., Haagsma, A.C., Chauhan, S., Heo, I., and Joo, C. (2016). TRBP ensures efficient Dicer processing of precursor microRNA in RNA-crowded environments. *Nat. Commun.* 7, 13694.

Fehring, G., Kraft, P., Pharoah, P.D., Eeles, R.A., Chatterjee, N., Schumacher, F.R., Schildkraut, J.M., Lindström, S., Brennan, P., Bickeböller, H., et al. (2016). Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer Res.* 76, 5103–5114.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J.W.W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *Eur. J. Cancer* 49, 1374–1403.

Fletcher, O., Johnson, N., Orr, N., Hosking, F.J., Gibson, L.J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C., et al. (2011). Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. *JNCI J. Natl. Cancer Inst.* 103, 425–435.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–

861.

Gaidatzis, D., Nimwegen, E. van, Hausser, J., Zavolan, M., Smit, A., Roskin, K., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E., et al. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinforma.* 2007 81 14, 708–715.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat. Struct. Mol. Biol.* 18, 1139–1146.

Ghanbari, M., de Vries, P.S., de Looper, H., Peters, M.J., Schurmann, C., Yaghootkar, H., Dörr, M., Frayling, T.M., Uitterlinden, A.G., Hofman, A., et al. (2014). A Genetic Variant in the Seed Region of miR-4513 Shows Pleiotropic Effects on Lipid and Glucose Homeostasis, Blood Pressure, and Coronary Artery Disease. *Hum. Mutat.* 35, 1524–1531.

Ghoussaini, M., Pharoah, P.D.P., and Easton, D.F. (2013). Inherited Genetic Susceptibility to Breast Cancer. *Am. J. Pathol.* 183, 1038–1051.

Giza, D.E., Vasilescu, C., and Calin, G.A. (2014). Key principles of miRNA involvement in human diseases. *Discov. (Craiova, Rom.)* 2, e34.

Glubb, D.M., Maranian, M.J., Michailidou, K., Pooley, K.A., Meyer, K.B., Kar, S., Carlebur, S., O'Reilly, M., Betts, J.A., Hillman, K.M., et al. (2015). Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* 96, 5–20.

Van Gool, I.C., Stelloo, E., Nout, R.A., Nijman, H.W., Edmondson, R.J., Church, D.N., MacKay, H.J., Leary, A., Powell, M.E., Mileskin, L., et al. (2016). Prognostic significance of L1CAM expression and its association with mutant p53 expression in high-risk endometrial cancer. *Mod. Pathol.* 29, 174–181.

Göring, H.H.H., Curran, J.E., Johnson, M.P., Dyer, T.D., Charlesworth, J., Cole, S.A., Jowett, J.B.M., Abraham, L.J., Rainwater, D.L., Comuzzie, A.G., et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39, 1208–1216.

Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engle, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Mol. Cell* 27, 91–105.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* 106, 23–34.

Ha, T.-Y. (2011). MicroRNAs in Human Diseases: From Cancer to Cardiovascular Disease. *Immune Netw.* 11, 135.

Haase, A.D., Jaskiewicz, L., Zhang, H., Lainé, S., Sack, R., Gatignol, A., and Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep.* 6,



961–967.

Hall, J., Lee, M., Newman, B., Morrow, J., Anderson, L., Huey, B., and King, M. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* (80-. ). 250, 1684–1689.

Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* 18, 3016–3027.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell* 125, 887–901.

Hatley, M.E., Patrick, D.M., Garcia, M.R., Richardson, J.A., Bassel-Duby, R., van Rooij, E., and Olson, E.N. (2010). Modulation of K-Ras-dependent lung tumorigenesis by MicroRNA-21. *Cancer Cell* 18, 282–293.

Haunsberger, S.J., Connolly, N.M.C., and Prehn, J.H.M. (2016). miRNAmeConverter: an R/bioconductor package for translating mature miRNA names to different miRBase versions. *Bioinformatics* 33, btw660.

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M. (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.* 19, 2009–2020.

He, N., Zheng, H., Li, P., Zhao, Y., Zhang, W., Song, F., and Chen, K. (2014). miR-485-5p Binding Site SNP rs8752 in HPGD Gene Is Associated with Breast Cancer Risk. *PLoS One* 9, e102093.

Hein, R., Maranian, M., Hopper, J.L., Kapuscinski, M.K., Southey, M.C., Park, D.J., Schmidt, M.K., Broeks, A., Hogervorst, F.B.L., Bueno-de-Mesquit, H.B., et al. (2012). Comparison of 6q25 Breast Cancer Hits from Asian and European Genome Wide Association Studies in the Breast Cancer Association Consortium (BCAC). *PLoS One* 7, e42380.

Hemminki, A., Markie, D., Tomlinson, I., Avizienyte, E., Roth, S., Loukola, A., Bignell, G., Warren, W., Aminoff, M., Höglund, P., et al. (1998). A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature* 391, 184–187.

Hoffman, A.E., Zheng, T., Yi, C., Leaderer, D., Weidhaas, J., Slack, F., Zhang, Y., Paranjape, T., and Zhu, Y. (2009). microRNA miR-196a-2 and breast cancer: A genetic and epigenetic association study and functional analysis. *Cancer Res.* 69, 5970–5977.

Hsu, J., Chiu, C.-M., Hsu, S.-D., Huang, W.-Y., Chien, C.-H., Lee, T.-Y., and Huang, H.-D. (2011). miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics* 12, 300.

Hu, S., Cui, G., Huang, J., Jiang, J., and Wang, D. (2016). An APOC3 3' UTR

variant associated with plasma triglycerides levels and coronary heart disease by creating a functional miR-4271 binding site. *Nat. Publ. Gr.* 1–10.

Huang, D., and Ovcharenko, I. (2015). Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res.* 43, 225–236.

Hulse, A.M., and Cai, J.J. (2013). Genetic variants contribute to gene expression variability in humans. *Genetics* 193, 95–108.

Hutvagner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science* 297, 2056–2060.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* 293, 834–838.

Iki, T., Yoshikawa, M., Nishikiori, M., Jaudal, M.C., Matsumoto-Yokoyama, E., Mitsuhashi, I., Meshi, T., and Ishikawa, M. (2010). In vitro assembly of plant RNA-induced silencing complexes facilitated by molecular chaperone HSP90. *Mol. Cell* 39, 282–291.

Iorio, M. V, Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., et al. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.* 65, 7065–7070.

Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T., and Tomari, Y. (2010). Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Mol. Cell* 39, 292–299.

Jiang, P., Li, Y., Poleshko, A., Medvedeva, V., Baulina, N., Zhang, Y., Zhou, Y., Slater, C.M., Pellegrin, T., Wasserman, J., et al. (2017). The Protein Encoded by the CCDC170 Breast Cancer Gene Functions to Organize the Golgi-Microtubule Network. *EBioMedicine* 22, 28–43.

Jiang, S., Zhang, H.W., Lu, M.H., He, X.H., Li, Y., Gu, H., Liu, M.F., and Wang, E.D. (2010). MicroRNA-155 functions as an oncomiR in breast cancer by targeting the suppressor of cytokine signaling 1 gene. *Cancer Res.* 70, 3119–3127.

Jin, H.Y., Gonzalez-martin, A., Miletic, A. V, Lai, M., and Knight, S. (2015). Transfection of microRNA Mimics Should Be Used with Caution. *Front. Genet.* 6, 1–23.

Jo, Y.K., Kim, S.C., Park, I.J., Park, S.J., Jin, D.-H., Hong, S.-W., Cho, D.-H., and Kim, J.C. (2012). Increased expression of ATG10 in colorectal cancer is associated with lymphovascular invasion and lymph node metastasis. *PLoS One* 7, e52705.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and De Bakker, P.I.W. (2008). SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938–2939.

Johnston, M., Geoffroy, M.-C., Sobala, A., Hay, R., and Hutvagner, G. (2010). HSP90 protein stabilizes unloaded argonaute complexes and microscopic P-bodies in human cells. *Mol. Biol. Cell* 21, 1462–1469.

Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53.

Kabirizadeh, S., Azadeh, M., Mirhosseini, M., Ghaedi, K., and Mesrian Tanha, H. (2016). The SNP rs3746444 within mir-499a is associated with breast cancer risk in Iranian population. *J. Cell. Immunother.* 2, 95–97.

Karginov, F. V., Cheloufi, S., Chong, M.M.W., Stark, A., Smith, A.D., and Hannon, G.J. (2010). Diverse Endonucleolytic Cleavage Sites in the Mammalian Transcriptome Depend upon MicroRNAs, Drosha, and Additional Nucleases. *Mol. Cell* 38, 781–788.

Kassie, F., Matise, I., Negia, M., Lahti, D., Pan, Y., Scherber, R., Upadhyaya, P., and Hecht, S.S. (2008). Combinations of N-Acetyl-S-(N-2-Phenethylthiocarbamoyl)-L-Cysteine and myo-inositol inhibit tobacco carcinogen-induced lung adenocarcinoma in mice. *Cancer Prev. Res. (Phila).* 1, 285–297.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* 15, 2654–2659.

Key, T.J., and Verkasalo, P.K. (1999). Endogenous hormones and the aetiology of breast cancer. *Breast Cancer Res.* 1, 18–21.

Key, T.J., Verkasalo, P.K., and Banks, E. (2001). Epidemiology of breast cancer. *Lancet. Oncol.* 2, 133–140.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell* 115, 209–216.

King, M.C., Rowell, S., and Love, S.M. (1993). Inherited breast and ovarian cancer. What are the risks? What are the choices? *Jama* 269, 1975–1980.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389.

Knight, S.W., and Bass, B.L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* (80-. ). 293, 2269–2271.

Koguchi, T., Tanikawa, C., Mori, J., Kojima, Y., and Matsuda, K. (2016). Regulation of myo-inositol biosynthesis by p53-ISYNA1 pathway. *Int. J. Oncol.* 48, 2415–2424.

Kontorovich, T., Levy, A., Korostishevsky, M., Nir, U., and Friedman, E. (2010). Single nucleotide polymorphisms in miRNA binding sites and miRNA genes as breast/ovarian cancer risk modifiers in Jewish high-risk women. *Int. J. Cancer*

127, 589–597.

Kozomara, A., and Griffiths-Jones, S. (2014). MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, 68–73.

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495–500.

Krüger, J., and Rehmsmeier, M. (2006). RNAhybrid: MicroRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* 34, 451–454.

Lagos-Quintana, M. (2001). Identification of Novel Genes Coding for Small Expressed RNAs. *Science* (80-. ). 294, 853–858.

Landi, D., Gemignani, F., Naccarati, A., Pardini, B., Vodicka, P., Vodickova, L., Hemminki, K., Novotny, J., Fo, A., and Landi, S. (2008). Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. 29, 579–584.

Lau, N.C. (2001). An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*. *Science* (80-. ). 294, 858–862.

Lee, A., and Park, J. (2016). Genetic variation rs7930 in the miR-4273-5p target site is associated with a risk of colorectal cancer. 6885–6895.

Lee, H.Y., and Doudna, J.A. (2012). TRBP alters human precursor microRNA processing in vitro. *RNA* 18, 2012–2019.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294, 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843–854.

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V.N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* 21, 4663–4670.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425, 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* 23, 4051–4060.

Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of Mammalian MicroRNA Targets. *Cell* 115, 787–798.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20.

Li, Y., and Kowdley, K. V. (2012). MicroRNAs in Common Human Diseases.

Genomics. Proteomics Bioinformatics 10, 246–253.

Li, M.J., Zhang, J., Liang, Q., Xuan, C., Wu, J., Jiang, P., Li, W., Zhu, Y., Wang, P., Fernandez, D., et al. (2017). Exploring genetic associations with ceRNA regulation in the human genome. *Nucleic Acids Res.* 45, 5653–5665.

Lim, L.P. (2003). Vertebrate MicroRNA Genes. *Science* (80-. ). 299, 1540–1540.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433, 769–773.

Liu, R., Maia, A.-T., Russell, R., Caldas, C., Ponder, B.A., and Ritchie, M.E. (2012). Allele-specific expression analysis methods for high-density SNP microarray data. *Bioinformatics* 28, 1102–1108.

Lo, H.S.S., Wang, Z., Hu, Y., Yang, H.H.H., Gere, S., Buetow, K.H.H., and Lee, M.P.P. (2003). Allelic variation in gene expression is common in the human genome. *Genome Res.* 13, 1855.

Long, D., Lee, R., Williams, P., Chan, C.Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* 14, 287–294.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* 435, 834–838.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear Export of MicroRNA. *Science* (80-. ). 303, 95–98.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901.

Mack, T., Hamilton, A., Press, M., Diep, A., and Rappaport, E. (2002). Heritable breast cancer in twins. *Br. J. Cancer* 87, 294–300.

Maia, A.-T., Antoniou, A.C., O'Reilly, M., Samarajiwa, S., Dunning, M., Kartsonaki, C., Chin, S.-F., Curtis, C.N., McGuffog, L., Domchek, S.M., et al. (2012). Effects of BRCA2 *cis*-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast Cancer Res.* 14, R63.

Majoros, W.H., Ohler, U., Baertsch, R., Barber, G., Bejerano, G., Clawson, H., Diekhans, M., Furey, T., Harte, R., Hsu, F., et al. (2007). Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics* 2007 81 34, D590-8.

Malkin, D., Li, F.P., Strong, L.C., Fraumeni, J.F., Nelson, C.E., Kim, D.H., Kassel, J., Gryka, M.A., Bischoff, F.Z., and Tainsky, M.A. (1990). Germ line p53

mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250, 1233–1238.

Mansfield, J.H., Harfe, B.D., Nissen, R., Obenaus, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A.E., Ruvkun, G., et al. (2004). MicroRNA-responsive “sensor” transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat. Genet.* 36, 1079–1083.

Mathew, R., Kongara, S., Beaudoin, B., Karp, C.M., Bray, K., Degenhardt, K., Chen, G., Jin, S., and White, E. (2007). Autophagy suppresses tumor progression by limiting chromosomal instability. *Genes Dev.* 21, 1367–1381.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2\*1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* 31, 55–59.

Meister, G., and Tuschli, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature* 431, 343–349.

Meyer, K.B., Maia, A.-T., O'Reilly, M., Teschendorff, A.E., Chin, S.-F., Caldas, C., and Ponder, B.A.J. (2008). Allele-Specific Up-Regulation of FGFR2 Increases Susceptibility to Breast Cancer. *PLoS Biol.* 6, e108.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* 45, 353–361.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* 47, 373–380.

Miyoshi, T., Takeuchi, A., Siomi, H., and Siomi, M.C. (2010). A direct role for Hsp90 in pre-RISC formation in *Drosophila*. *Nat. Struct. Mol. Biol.* 17, 1024–1026.

Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A., Schadt, E.E., Sachs, A., et al. (2004). Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–1105.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747.

Morselli, E., Galluzzi, L., Kepp, O., Vicencio, J.M., Criollo, A., Maiuri, M.C., and

- Kroemer, G. (2009). Anti- and pro-tumor functions of autophagy. *Biochim. Biophys. Acta - Mol. Cell Res.* 1793, 1524–1532.
- Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* 16, 720–728.
- Nakaoka, H.J., Tanei, Z., Hara, T., Weng, J.S., Kanamori, A., Hayashi, T., Sato, H., Orimo, A., Otsuji, K., Tada, K., et al. (2017). Mint3-mediated L1CAM expression in fibroblasts promotes cancer cell proliferation via integrin  $\alpha 5 \beta 1$  and tumour growth. *Oncogenesis* 6, e334.
- Narod, S.A., Feunteun, J., Lynch, H.T., Watson, P., Conway, T., Lynch, J., and Lenoir, G.M. (1991). Familial breast-ovarian cancer locus on chromosome 17q12-q23. *Lancet (London, England)* 338, 82–83.
- Nemoto, T., Tanida, I., Tanida-Miyake, E., Minematsu-Ikeguchi, N., Yokota, M., Ohsumi, M., Ueno, T., and Kominami, E. (2003). The mouse APG10 homologue, an E2-like enzyme for Apg12p conjugation, facilitates MAP-LC3 modification. *J. Biol. Chem.* 278, 39517–39526.
- Newman, B., Austin, M.A., Lee, M., and King, M.C. (1988). Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc. Natl. Acad. Sci.* 85, 3044–3048.
- Nicoloso, M.S., Sun, H., Spizzo, R., Kim, H., Wickramasinghe, P., Shimizu, M., Wojcik, S.E., Ferdin, J., Kunej, T., Xiao, L., et al. (2010). Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Res.* 70, 2789–2798.
- Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C.B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 13, 1894–1910.
- Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genet.* 11.
- Panwar, B., Omenn, G.S., and Guan, Y. (2017). miRmine: A Database of Human miRNA Expression Profiles. *Bioinformatics* 44, btx019.
- Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* 41, W169–W173.
- Park, J.-E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., and Kim, V.N. (2011). Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature* 475, 201–205.
- Pasquinelli, a E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic

regulatory RNA. *Nature* 408, 86–89.

Pastinen, T., and Hudson, T.J. (2004). Cis-Acting Regulatory Variation in the Human Genome. *Science* (80-. ). 306, 647–650.

Patel, K.J., Yu, V.P., Lee, H., Corcoran, A., Thistlethwaite, F.C., Evans, M.J., Colledge, W.H., Friedman, L.S., Ponder, B.A., and Venkitaraman, A.R. (1998). Involvement of Brca2 in DNA repair. *Mol. Cell* 1, 347–357.

Philips, S., Richter, A., Oesterreich, S., Rae, J.M., Flockhart, D.A., Perumal, N.B., and Skaar, T.C. (2012). Functional Characterization of a Genetic Polymorphism in the Promoter of the ESR2 Gene. *Horm. Cancer* 3, 37–43.

Pierce, B.L., Tong, L., Chen, L.S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.-J., Franke, L., et al. (2014). Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet.* 10, e1004818.

Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., Stefansson, K., and Spielman, R. (2011). Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLoS Genet.* 7, e1001317.

R Core Team (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., et al. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* 39, 165–167.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., and Hatzigeorgiou, A.G. (2012). Functional microRNA targets in protein coding sequences. *Bioinformatics* 28, 771–776.

Rehmsmeier, M., Steffen, P., Höchsmann, M., Giegerich, R., and Ho, M. (2004). Fast and effective prediction of microRNA / target duplexes. *Spring* 1507–1517.

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., North, B., Jayatilake, H., Barfoot, R., Spanova, K., et al. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* 38, 873–875.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513–520.

Robins, H., and Press, W.H. (2005). Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15557–15562.

Robins, H., Li, Y., and Padgett, R.W. (2005). Incorporating structure to predict microRNA targets. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4006–4009.

Rosenberg, N.A., and VanLiere, J.M. (2009). Replication of genetic associations as pseudoreplication due to shared genealogy. *Genet. Epidemiol.* 33, 479–487.



- Rowell, S., Newman, B., Boyd, J., and King, M.C. (1994). Inherited predisposition to breast and ovarian cancer. *Am. J. Hum. Genet.* 55, 861–865.
- Russo, J., and Russo, I.H. (2006). The role of estrogen in the initiation of breast cancer. *J. Steroid Biochem. Mol. Biol.* 102, 89–96.
- Ruvkun, G., Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., and Horvitz, H.R. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901–906.
- Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302.
- Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. *Science* (80-. ). 346, 608–613.
- Schmidt, M., and Finley, D. (2014). Regulation of proteasome activity in health and disease. *Biochim. Biophys. Acta - Mol. Cell Res.* 1843, 13–25.
- Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115, 199–208.
- Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., et al. (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* 38, 1239–1241.
- Selcuklu, S.D., Donoghue, M.T.A., and Spillane, C. (2009). miR-21 as a key regulator of oncogenic processes. *Biochem. Soc. Trans.* 37, 918–925.
- Sempere, L.F., Christensen, M., Silahatoglu, A., Bak, M., Heath, C. V, Schwartz, G., Wells, W., Kauppinen, S., and Cole, C.N. (2007). Altered MicroRNA expression confined to specific epithelial cell subpopulations in breast cancer. *Cancer Res.* 67, 11612–11620.
- Sharan, S.K., Morimatsu, M., Albrecht, U., Lim, D.-S., Regel, E., Dinh, C., Sands, A., Eichele, G., Hasty, P., and Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature* 386, 804–810.
- Shin, C., Nam, J.-W., Farh, K.K.-H., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing. *Mol. Cell* 38, 789–802.
- Si, M.-L., Zhu, S., Wu, H., Lu, Z., Wu, F., and Mo, Y.-Y. (2007). miR-21-mediated tumor growth. *Oncogene* 26, 2799–2803.
- Sinnett, D., Beaulieu, P., Bélanger, H., Lefebvre, J.-F., Langlois, S., Théberge, M.-C., Drouin, S., Zotti, C., Hudson, T.J., and Labuda, D. (2006). Detection and characterization of DNA variants in the promoter regions of hundreds of human

disease candidate genes. *Genomics* 87, 704–710.

Smith, R.A., Jedlinski, D.J., Gabrovskaa, P.N., Weinstein, S.R., Haupt, L., and Griffiths, L.R. (2012). A genetic variant located in miR-423 is associated with reduced breast cancer risk. *Cancer Genomics Proteomics* 9, 115–118.

Song, C., Chen, G.K., Millikan, R.C., Ambrosone, C.B., John, E.M., Bernstein, L., Zheng, W., Hu, J.J., Ziegler, R.G., Nyante, S., et al. (2013). A Genome-Wide Scan for Breast Cancer Risk Haplotypes among African American Women. *PLoS One* 8, 32118–32119.

Soule, H.D., Vazquez, J., Long, A., Albert, S., and Brennan, M. (1973). A human cell line from a pleural effusion derived from a breast carcinoma. *J. Natl. Cancer Inst.* 51, 1409–1416.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007a). Population genomics of human gene expression. *Nat. Genet.* 39, 1217–1224.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007b). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science* (80-. ). 315, 848–853.

Sudarsanam, S., and Johnson, D.E. (2010). Functional consequences of mTOR inhibition. *Curr. Opin. Drug Discov. Devel.* 13, 31–40.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D.A., Sommer, S.S., and Rossi, J.J. (2009). SNPs in human miRNA genes affect biogenesis and function. *RNA* 15, 1640–1651.

Swift, M., Sholman, L., Perry, M., and Chase, C. (1976). Malignant Neoplasms in the Families of Patients with Ataxia-Telangiectasia. *Cancer Res.* 36.

Tafer, H., Ameres, S.L., Obernosterer, G., Gebeshuber, C.A., Schroeder, R., Martinez, J., and Hofacker, I.L. (2008). The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.* 26, 578–583.

Tangen, I.L., Kopperud, R.K., Visser, N.C., Staff, A.C., Tingulstad, S., Marcickiewicz, J., Amant, F., Bjørge, L., Pijnenborg, J.M., Salvesen, H.B., et al. (2017). Expression of L1CAM in curettage or high L1CAM level in preoperative blood samples predicts lymph node metastases and poor outcome in endometrial cancer patients. *Br. J. Cancer* 117, 840–847.

Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., Byrd, P., Taylor, M., and Easton, D.F. (2005). Cancer Risks and Mortality in Heterozygous ATM Mutation Carriers. *JNCI J. Natl. Cancer Inst.* 97, 813–822.

Thorisson, G.A., Smith, A. V, Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web site. *Genome Res.* 15, 1592–1593.

Toraih, E.A., Hussein, M.H., Al Ageeli, E., Riad, E., AbdAllah, N.B., Helal, G.M., and Fawzy, M.S. (2017). Structure and functional impact of seed region variant in MIR-499 gene family in bronchial asthma. *Respir. Res.* 18, 169.

Travis, R.C., and Key, T.J. (2003). Oestrogen exposure and breast cancer risk. *Breast Cancer Res.* 5, 239.

VanLiere, J.M., and Rosenberg, N.A. (2008). Mathematical properties of the  $r^2$  measure of linkage disequilibrium. *Theor. Popul. Biol.* 74, 130–137.

Veeraraghavan, J., Tan, Y., Cao, X.-X., Kim, J.A., Wang, X., Chamness, G.C., Maiti, S.N., Cooper, L.J.N., Edwards, D.P., Contreras, A., et al. (2014). Recurrent ESR1–CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat. Commun.* 5, ncomms5577.

Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet.* 4, e1000214.

Wang, H., Hang, C., Ou, X.-L., Nie, J.-S., Ding, Y.-T., Xue, S.-G., Gao, H., and Zhu, J.-X. (2016). MiR-145 functions as a tumor suppressor via regulating angiopoietin-2 in pancreatic cancer cells. *Cancer Cell Int.* 16, 65.

Wattenberg, L.W., and Estensen, R.D. (1996). Chemopreventive effects of myo-inositol and dexamethasone on benzo[a]pyrene and 4-(methylnitrosoamino)-1-(3-pyridyl)-1-butanone-induced pulmonary carcinogenesis in female A/J mice. *Cancer Res.* 56, 5132–5135.

Wen, J., Parker, B.J., Jacobsen, A., and Krogh, A. (2011). MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA* 17, 820–834.

Wheeler, G., Ntounia-Fousara, S., Granda, B., Rathjen, T., and Dalmay, T. (2006). Identification of new central nervous system specific mouse microRNAs. *FEBS Lett.* 580, 2195–2200.

Wickham, H. (2010). *ggplot2: Elegant Graphics for Data Analysis*.

Williams, R.B.H., Chan, E.K.F., Cowley, M.J., and Little, P.F.R. (2007). The influence of genetic variation on gene expression. *Genome Res.* 17, 1707–1716.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., et al. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* (80-. ). 265, 2088–2090.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins,

- N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378, 789–792.
- Wu, H., Zhu, S., and Mo, Y.-Y. (2009). Suppression of cell growth and invasion by miR-205 in breast cancer. *Cell Res.* 19, 439–448.
- Wynendaele, J., Böhnke, A., Leucci, E., Nielsen, S.J., Lambert, I., Hammer, S., Sbrzesny, N., Kubitz, D., Wolf, A., Gradhand, E., et al. (2010). An illegitimate microRNA target site within the 3' UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Res.* 70, 9641–9649.
- Xavier, J., Russell, R., Almeida, B.P., Rosli, N., Rocha, C., Samarajiwa, S., Chin, S.-F., Caldas, C., Ponder, B.A., and Maia, A.-T. (2016). Abstract A31: Integrative differential allelic expression analysis efficiently reveals the biology underlying risk to breast cancer. *Mol. Cancer Res.* 14, A31–A31.
- Xia, B., Sheng, Q., Nakanishi, K., Ohashi, A., Wu, J., Christ, N., Liu, X., Jasin, M., Couch, F.J., and Livingston, D.M. (2006). Control of BRCA2 Cellular and Clinical Functions by a Nuclear Partner, PALB2. *Mol. Cell* 22, 719–729.
- Xu, L., Yu, J., Wang, Z., Zhu, Q., Wang, W., and Lan, Q. (2017). miR-543 functions as a tumor suppressor in glioma in vitro and in vivo. *Oncol. Rep.* 38, 725–734.
- Xu, X., Wagner, K.-U., Larson, D., Weaver, Z., Li, C., Ried, T., Hennighausen, L., Wynshaw-Boris, A., and Deng, C.-X. (1999). Conditional mutation of Brca1 in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation. *Nat. Genet.* 22, 37–43.
- Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., and Kinzler, K.W. (2002a). Allelic variation in human gene expression. *Science* 297, 1143.
- Yan, H., Dobbie, Z., Gruber, S.B., Markowitz, S., Romans, K., Giardiello, F.M., Kinzler, K.W., and Vogelstein, B. (2002b). Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.* 30, 25–26.
- Yan, L.-X., Huang, X.-F., Shao, Q., Huang, M.-Y., Deng, L., Wu, Q.-L., Zeng, Y.-X., and Shao, J.-Y. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *RNA* 14, 2348–2360.
- Yang, C.H., Pfeffer, S.R., Sims, M., Yue, J., Wang, Y., Ling, V.G., Paulus, E., Davidoff, A.M., and Pfeffer, L.M. (2015). The oncogenic microRNA-21 inhibits the tumor suppressive activity of FBXO11 to promote tumorigenesis. *J. Biol. Chem.* 290, 6037–6046.
- Yang, F., Wang, J., The GTEx Consortium, Pierce, B.L., and Chen, L.S. (2016). Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *bioRxiv*.
- Yang, J.-S., Phillips, M.D., Betel, D., Mu, P., Ventura, A., Siepel, A.C., Chen, K.C., and Lai, E.C. (2011). Widespread regulatory activity of vertebrate microRNA\* species. *RNA* 17, 312–326.
- Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D.,

Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. *Nucleic Acids Res.* 44, D710–D716.

Yekta, S., Shih, I.-H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304, 594–596.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* 17, 3011–3016.

Zhao, L., Yu, H., Yi, S., Peng, X., Su, P., Xiao, Z., Liu, R., Tang, A., Li, X., Liu, F., et al. (2016). The tumor suppressor miR-138-5p targets PD-L1 in colorectal cancer. *Oncotarget* 7, 45370–45384.

Zhong, X., Coukos, G., and Zhang, L. (2012). miRNAs in human cancer. *Methods Mol. Biol.* 822, 295–306.

Zhu, J., Zheng, Z., Wang, J., Sun, J., Wang, P., Cheng, X., Fu, L., Zhang, L., Wang, Z., and Li, Z. (2014). Different miRNA expression profiles between human breast cancer tumors and serum. *Front. Genet.* 5, 1–7.

Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Bin Li, B., Zheng, Y., and Shi, L. (2016). Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Sci. Rep.* 6, 28400.

(2007). microRNAs as oncogenes and tumor suppressors. *Dev. Biol.* 302, 1–12.

References are according to submission rules of Cell.

# ANNEXES

## Annex A

### Annex A. List of GWAS-significant SNPs.

rs12405132	rs2046210	rs2823093	rs17530068	rs115392158
rs12048493	rs6964587	rs17879961	rs9628987	rs11200014
rs72755295	rs720475	rs132390	rs6797852	rs7931342
rs6796502	rs9693444	rs6001930	rs17051310	rs148883465
rs13162653	rs6472903	rs13393577	rs4455437	rs11168936
rs2012709	rs2943559	rs10871290	rs6835704	rs73110464
rs7707921	rs13281615	rs10822013	rs3806872	rs11571818
rs9257408	rs11780156	rs1219648	rs11613298	rs11844632
rs4593472	rs1011970	rs7716600	rs6788895	rs141752671
rs13365225	rs10759243	rs4784227	rs3750817	rs11065979
rs13267382	rs865686	rs653465	rs903263	rs56084662
rs11627032	rs2380205	rs2229882	rs1314913	rs4951011
rs745570	rs7072776	rs2392780	rs16882214	rs10474352
rs6507583	rs11814448	rs2912774	rs12906542	rs2290203
chr17:29230520:D	rs10995190	rs537626	rs1078806	rs11082321
rs616488	rs704010	rs4784223	rs2305016	rs2075555
rs11552449	rs7904519	rs16886181	rs17435444	rs6556756
rs11249433	rs11199914	rs16886448	rs4414128	rs1154865
rs6678914	rs2981579	rs16886397	rs765899	rs1978503
rs4245739	rs3817198	rs3822625	rs140068132	rs1926657
rs12710696	rs3903072	rs16886364	rs2912780	rs10263639
rs4849887	rs78540526	rs16886113	rs2193094	rs10490113
rs2016394	rs554219	rs1017226	rs13025833	rs2290854
rs1550623	rs75915166	rs7726354	rs12118297	rs11196174
rs1045485	rs11820646	rs12655019	rs16992204	rs765855
rs13387042	rs12422552	rs16886034	rs12628403	rs8100241
rs16857609	rs10771399	rs614367	rs2974935	rs9383938
rs6762644	rs17356907	rs2180341	rs1057941	rs2284378
rs4973768	rs1292011	rs2981582	rs11119608	rs7535752
rs12493607	rs11571833	rs16886165	rs75316749	rs1810320

rs1053338	rs2236007	rs2981575	rs115707823	rs13116936
rs9790517	rs2588809	rs4415084	rs147527678	rs921551
rs6828523	rs999737	rs1562430	rs3184504	rs16875333
rs10069690	rs941764	rs3112612	rs12601991	rs2230754
rs7726159	rs3803662	rs3734805	rs4808075	rs12711517
rs2736108	rs17817449	rs10510102	rs11907546	rs2289731
rs10941679	rs11075995	rs1092913	rs56404467	rs17141741
rs889312	rs13329835	rs9485372	rs2300206	rs2386661
rs10472076	rs6504950	rs9383951	rs186507655	rs2842346
rs1353747	rs527616	rs7107217	rs2075570	rs737387
rs1432679	rs1436904	rs909116	rs481519	rs2842347
rs11242675	rs8170	rs3757318	rs7679673	rs10132579
rs204247	rs2363956	rs4322600	rs7725218	rs757369
rs17529111	rs4808801	rs2981578	rs1862626	
rs12662670	rs3760982	rs12922061	rs147680653	

## Annex B

**Annex B. 3'UTR-located SNPs.** 64 unique SNPs are listed. Underlined are GWAS-significant SNPs which were not genotyped in the 1000 Genomes Project and therefore proxy SNPs were not retrieved. In bold are GWAS-significant SNPs also located in the 3'UTR. Both bold and underlined are genes without evidence of significant differential allelic expression. SNPs which were not analysed by the TargetScan algorithm, because their genes were not in the provided dataset of 3'UTR sequences, are highlighted in light grey. Highlighted in dark grey are SNPs which did not pass the location control (see subchapter 4.4.1) and thus, were not analysed by TargetScan.

GWAS SNP	r <sup>2</sup>	SNP	Chr	Allele	Ancestral allele	Gene
rs903263	0.965	rs1057738	1	A/C	A	PRKACB
rs903263	0.932	rs10782824	1	T/A	A	PRKACB
rs2290854	1	rs10900596	1	T/C	C	MDM4
rs4245739	0.861	rs10900596	1	T/C	C	MDM4
rs2290854	1	rs10900597	1	C/T	C	MDM4
rs4245739	0.861	rs10900597	1	C/T	C	MDM4
rs12405132	0.851	rs12123298	1	G/A/C	G	RNF115
rs12405132	0.961	rs17352469	1	T/C	T	RNF115
rs12405132	0.851	rs17354678	1	T/C	T	RNF115

# The role of miRNA-mediated *cis*-regulation in breast cancer susceptibility

Ana Catarina Jacinta Fernandes © 2017

rs12405132	0.961	rs1778523	1	G/C	G	<b>CD160</b>
rs12405132	0.961	rs2231375	1	C/T	C	<b>CD160</b>
rs903263	0.901	rs3768258	1	A/G	A	PRKACB
rs2290854	0.964	rs4245738	1	C/T	T	MDM4
rs4245739	0.821	rs4245738	1	C/T	T	MDM4
rs2290854	0.861	rs4245739	1	C/A	A	MDM4
rs4245739	1	<b>rs4245739</b>	1	C/A	A	MDM4
rs1053338	0.837	rs1046025	3	C/T	T	<b>PSMD6</b>
rs4973768	1	rs1051545	3	T/C	C	SLC4A7
rs653465	0.839	rs1051545	3	T/C	C	SLC4A7
rs1053338	0.895	rs3733126	3	C/T	T	ATXN7
rs4973768	1	<b>rs4973768</b>	3	C/T	T	SLC4A7
rs653465	0.839	rs4973768	3	C/T	T	SLC4A7
rs7707921	1	rs1019806	5	G/A	A	ATG10
rs12655019	0.92	rs12654125	5	G/A	G	MIER3
rs1092913	0.908	rs1287599	5	C/A/G	A	MARCH6
rs1092913	0.908	rs1287600	5	C/T	C	MARCH6
rs1092913	0.908	rs1287601	5	T/G	T	MARCH6
rs12655019	0.92	rs1466010	5	A/G	A	SETD9
rs12655019	0.92	rs16886496	5	T/C	T	MIER3
rs1092913	0.908	rs2589668	5	A/G	A	MARCH6
rs12655019	0.92	rs3756586	5	A/G	G	MIER3
rs3806872	1	<b>rs3806872</b>	5	C/A/T	C	ADAMTS16
rs7707921	1	rs6884232	5	G/A	A	ATG10
rs7707921	1	rs1019806	5	A/G	A	ATG10
rs7707921	0.881	rs73136782	5	T/G	T	ATG10
rs12662670	0.892	rs3734805	6	A/C	A	CCDC170
rs3734805	1	<b>rs3734805</b>	6	A/C	A	CCDC170
rs2046210	0.821	rs3734806	6	G/A	G	CCDC170
rs2046210	0.821	rs3757322	6	T/G	T	CCDC170
rs2180341	1	rs9321073	6	C/T	T	RNF146
rs12662670	0.892	rs9383589	6	A/G	A	CCDC170
rs3734805	1	rs9383589	6	A/G	A	CCDC170
rs12662670	0.892	rs9383935	6	C/T	C	CCDC170
rs3734805	1	rs9383935	6	C/T	C	CCDC170
rs6964587	1	rs10225885	7	A/G	A	AKAP9
rs6964587	1	rs10225892	7	A/G	G	AKAP9
rs6964587	1	rs28584017	7	G/A	G	AKAP9
rs6964587	1	rs4265	7	C/T	C	AKAP9



rs6964587	1	rs55745934	7	T/C	T	AKAP9
rs6964587	0.934	rs7793861	7	C/G	G	<b>CYP51A1</b>
rs2386661	0.826	rs1132293	10	C/T	C	ASB13
rs4414128	0.832	rs1573	10	A/G	A	ASB13
rs4414128	0.832	rs2386648	10	T/A	A	ASB13
rs4414128	0.832	rs2386649	10	C/T	C	ASB13
rs3903072	0.837	rs633800	11	G/A	G	EFEMP2
<u>rs56084662</u>	1	<b>rs56084662</b>	13	G/A	G	FRY
rs6504950	0.83	rs17817901	17	A/G	A	COX11
rs6504950	0.83	rs17817901	17	A/G	A	TOM1L1
rs6504950	0.83	rs1802212	17	A/C	A	TOM1L1
rs6504950	0.83	rs1802212	17	A/C	A	COX11
rs6504950	0.83	rs3087650	17	G/A	G	COX11
rs4808801	1	rs10405636	19	A/C	C	SSBP4
rs8170	1	rs10425939	19	C/T	C	ANKLE1
rs4808801	0.897	rs1043327	19	A/G	G	ELL
rs4808801	1	rs10442	19	A/C/G/T	C	SSBP4
rs4808801	1	rs10442	19	A/C/G/T	C	ISYNA1
rs8170	0.859	rs11540855	19	A/G/T	A	ABHD8
rs4808801	1	rs2303697	19	T/C	C	ISYNA1
rs2363956	1	<b>rs2363956</b>	19	T/G	T	ANKLE1
rs8100241	1	rs2363956	19	T/G	T	ANKLE1
rs4808801	0.965	rs2385088	19	A/G	G	ISYNA1
rs4808075	1	rs4808616	19	C/A	C	ABHD8
rs2363956	1	rs8100241	19	G/A	G	ENSG00000269307
rs8100241	1	<b>rs8100241</b>	19	G/A	G	ENSG00000269307
rs2363956	1	rs8108174	19	T/A	T	ENSG00000269307
rs2363956	1	rs8108174	19	T/A	T	ANKLE1
rs8100241	1	rs8108174	19	T/A	T	ENSG00000269307
rs8100241	1	rs8108174	19	T/A	T	ANKLE1
rs8170	1	<b>rs8170</b>	19	G/A	G	BABAM1
rs2284378	1	rs6119447	20	A/G	G	RALY
rs2284378	1	rs8123521	20	A/C	C	RALY
<u>rs17879961</u>	1	<b>rs17879961</b>	22	A/C/G	A	CHEK2

## Annex C

**Annex C. 54 unique SNPs located in the 5'UTR and/or CDS are listed.** Underlined are GWAS-significant SNPs which were not genotyped in the 1000 Genomes Project and therefore proxy SNPs were not retrieved. In bold are GWAS-significant SNPs also located either in the 5'UTR or CDS of PCGs. Highlighted in grey are SNPs which also are located in 3'UTRs, although in a different gene.

GWAS SNP	r <sup>2</sup>	SNP	Chr	Allele	Ancestral allele	Gene	Location
rs2289731	0.897	rs1058619	1	A/G	G	SLC45A1	CDS
rs11552449	1	<b>rs11552449</b>	1	C/G/T	C	DCLRE1B	CDS
rs2289731	1	<b>rs2289731</b>	1	G/A	G	SLC45A1	CDS
rs11552449	1	rs3761936	1	T/C	T	DCLRE1B	CDS
rs4951011	1	<b>rs4951011</b>	1	A/G	A	ZBED6	5'UTR
rs4951011	1	rs4951011	1	A/G	A	ZC3H11A	5'UTR
rs9628987	0.925	rs56125713	1	G/C/T	G	SLC45A1	CDS
rs7535752	1	<b>rs7535752</b>	1	G/A/T	G	SLC45A1	CDS
rs9628987	0.81	rs7535752	1	G/A/T	G	SLC45A1	CDS
rs1045485	1	<b>rs1045485</b>	2	G/C	G	CASP8	CDS
rs1045485	1	rs17468277	2	C/T	C	ALS2CR12	CDS
rs2016394	0.905	rs743605	2	C/T	C	DLX2	5'UTR
rs1053338	1	<b>rs1053338</b>	3	A/G	A	ATXN7	CDS
rs481519	0.875	rs11129270	3	A/G	G	NEK10	5'UTR
rs653465	0.905	rs11129270	3	A/G	G	NEK10	5'UTR
rs481519	0.935	rs3213930	3	G/C/T	C	NEK10	CDS
rs4973768	0.81	rs3213930	3	G/C/T	C	NEK10	CDS
rs653465	0.967	rs3213930	3	G/C/T	C	NEK10	CDS
rs1053338	1	rs3733121	3	C/G	C	ATXN7	5'UTR
rs17051310	0.826	rs17051298	4	T/C/G	T	TNIP3	CDS
rs12655019	1	rs10042998	5	A/G	A	SETD9	5'UTR
rs1432679	0.899	rs1368298	5	A/G	G	EBF1	CDS
rs16886034	0.826	rs2229882	5	C/T	C	MAP3K1	CDS
rs16886113	1	rs2229882	5	C/T	C	MAP3K1	CDS
rs2229882	1	<b>rs2229882</b>	5	C/T	C	MAP3K1	CDS
rs2736108	0.933	rs2736098	5	C/T	C	TERT	CDS
rs1017226	1	rs3822625	5	A/G	A	MAP3K1	CDS
rs16886364	1	rs3822625	5	A/G	A	MAP3K1	CDS
rs16886397	1	rs3822625	5	A/G	A	MAP3K1	CDS
rs16886448	1	rs3822625	5	A/G	A	MAP3K1	CDS
rs3822625	1	<b>rs3822625</b>	5	A/G	A	MAP3K1	CDS

# The role of miRNA-mediated *cis*-regulation in breast cancer susceptibility

Ana Catarina Jacinta Fernandes © 2017

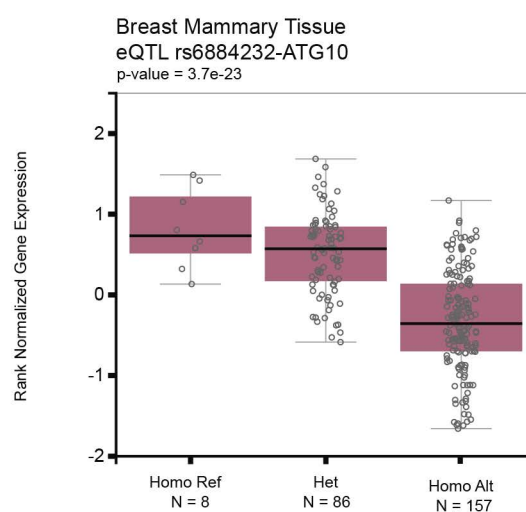
rs2180341	0.96	rs3756996	6	G/A/C	G	ECHDC1	5'UTR
rs2046210	0.821	rs6929137	6	G/A	G	CCDC170	CDS
rs2180341	0.924	rs877661	6	T/G	G	RNF146	5'UTR
rs6964587	1	rs10228334	7	C/T	T	AKAP9	CDS
rs6964587	0.967	rs10231350	7	G/T	T	LRRD1	5'UTR
rs6964587	1	rs10236397	7	C/T	C	AKAP9	CDS
rs6964587	1	rs1063243	7	A/C	C	AKAP9	CDS
rs6964587	1	rs13245393	7	A/G	G	AKAP9	CDS
rs6964587	1	rs28927678	7	C/T	C	AKAP9	CDS
rs6964587	1	rs4727267	7	G/C	G	AKAP9	5'UTR
rs6964587	0.967	rs6465353	7	T/G	G	LRRD1	CDS
rs6964587	1	rs6960867	7	A/G	A	AKAP9	CDS
rs6964587	1	<b>rs6964587</b>	7	G/T	T	AKAP9	CDS
rs6964587	1	rs7797834	7	A/G	A	CYP51A1	CDS
rs13365225	0.941	rs16885577	8	A/G	A	KCNU1	CDS
rs3903072	0.837	rs633800	11	G/A	G	EFEMP2	CDS,3'UTR
rs2230754	1	<b>rs2230754</b>	12	C/T	C	PLXNC1	CDS
rs3184504	1	<b>rs3184504</b>	12	T/C	C	SH2B3	CDS
rs1926657	0.932	rs899494	13	A/G	G	ABCC4	CDS
rs2290203	0.938	rs2301826	15	C/T	T	PRC1	CDS
rs6504950	0.889	rs1156287	17	G/A	A	STXBP4	CDS
rs6504950	0.83	rs1802212	17	A/C	A	COX11	CDS, 3'UTR
rs745570	0.902	rs4889891	17	C/A	A	CBX8	CDS
rs745570	0.902	rs9905914	17	G/A	A	CBX8	CDS
rs4808801	1	rs10405636	19	A/C	C	SSBP4	5'UTR, CDS, 3'UTR
rs8170	1	rs10424178	19	C/T	C	BABAM1	5'UTR
rs8170	1	rs10425939	19	C/T	C	ANKLE1	CDS,3'UTR
rs4808801	1	rs2303697	19	T/C	C	ISYNA1	CDS, 3'UTR
rs2363956	1	<b>rs2363956</b>	19	T/G	T	ANKLE1	CDS, 3'UTR
rs8100241	1	rs2363956	19	T/G	T	ANKLE1	CDS, 3'UTR
rs8170	1	rs73509996	19	T/G	T	USHBP1	5'UTR
rs2363956	1	rs8100241	19	G/A	G	ANKLE1	CDS
rs8100241	1	<b>rs8100241</b>	19	G/A	G	ANKLE1	CDS
rs2363956	1	rs8108174	19	T/A	T	USHBP1	5'UTR
rs2363956	1	rs8108174	19	T/A	T	ANKLE1	CDS, 3'UTR
rs8100241	1	rs8108174	19	T/A	T	USHBP1	5'UTR
rs8100241	1	rs8108174	19	T/A	T	ANKLE1	CDS, 3'UTR

rs8170	1	<b>rs8170</b>	19	G/A	G	BABAM1	CDS, 3'UTR
<u>rs17879961</u>	1	rs17879961	22	A/C/G	A	CHEK2	5'UTR, CDS, 3'UTR

## Annex D

**Annex D.** TargetScan analysis pipeline. *Online version only.*

## Annex E



**Annex E. eQTL association of rs6884232 with ATG10 expression levels in breast tissue.** Homo Ref are GG homozygotes, Het are GA heterozygotes and Homo Alt are AA homozygotes.